Automated Feedback Generation for Open-Ended Questions: Insights from Fine-Tuned LLMs

Elisabetta Mazzullo MAZZULLO@UALBERTA.CA
Measurement, Evaluation, and Data Science
University of Alberta, Edmonton, AB T6G 2G5 CANADA

Okan Bulut BULUT@UALBERTA.CA
Centre for Research in Applied Measurement and Evaluation
University of Alberta, Edmonton, AB T6G 2G5 CANADA

Abstract

Timely, personalized, and actionable feedback is essential for effective learning but challenging to deliver at scale. Automated feedback generation (AFG) using large language models (LLMs) can be a promising solution to address this challenge. While existing studies using out-of-the-box LLMs and prompting strategies have shown promise, there is room for improvement. This study investigates the fine-tuning of OpenAI's GPT-3.5-turbo for AFG. We developed feedback for open-ended situational judgment questions, and this small set of hand-crafted feedback examples was used to fine-tune the pre-trained LLM using specific prompting strategies. Our evaluation, conducted by independent judges and test experts, found that the feedback generated by our fine-tuned GPT-3.5-turbo model achieved high user satisfaction (84.8%) and met key structural quality criteria (72.9%). Also, the model generalized effectively across different items, providing feedback consistent with instructions, regardless of the respondent's performance level, English proficiency, or student status. However, some feedback statements still contained linguistic errors, lacked focused suggestions, or seemed generic. We discuss potential solutions to these issues, along with implications for developing LLM-supported AFG systems and their adoption in high-stakes settings.

Keywords: automated feedback generation, open-ended question, LLM, fine-tuning, GPT.

1. Introduction

Extensive research underscores the critical role of feedback in education (Hattie and Timperley, 2007) and advocates for a shift from assessment of learning to assessment for learning (Wiliam, 2011). The literature strongly recommends that educators provide feedback that is timely, personalized, and detailed (Hattie and Timperley, 2007), encourages active student engagement through dialogue (Carless, 2016), and offers actionable suggestions for improvement (Sadler, 2010b). However, generating feedback that aligns with these criteria imposes significant demands on educators (Boud and Dawson, 2023), who are already burdened with substantial workloads and the risk of burnout (Jomuad et al., 2021). Moreover, delivering such high-quality feedback is often impractical in the context of large-scale assessments or large classrooms, such as those in massive open online courses (MOOCs). Yet, feedback remains essential in these settings, particularly when assessments have significant consequences for learners or when online education limits access to instructor and peer support.

Leveraging their advanced language comprehension and generative capabilities, large language models (LLMs) present a promising solution for delivering timely and personalized feedback at scale. Moreover, due to their instruction-following abilities, LLMs enable educators to retain a degree of control in the feedback process by providing specific directions or examples to guide the model toward the desired output (e.g., Meyer et al., 2024). Although early research on using LLMs for automated feedback generation (AFG) has yielded encouraging results (Matelsky et al., 2023), the field remains nascent and rapidly evolving, with many aspects still unexplored. Notably, most studies have focused on large proprietary LLMs like OpenAI's GPT models. However, smaller open-source LLMs, such as Llama and Mistral, are also available and could offer a more accessible and cost-effective solution for researchers and small organizations while still delivering high performance. Also, researchers often utilize GPT in its ready-to-use chat version (i.e., ChatGPT), primarily relying on prompting techniques to shape the feedback output. This approach highlights the potential of LLMs but also underscores the need for further exploration into alternative models and more sophisticated methods for feedback generation.

Recently developed fine-tuning techniques provide a cost-effective way of adapting pre-trained LLMs to specific tasks (Pu et al., 2023), often without extensive tuning datasets (Jha et al., 2023). Fine-tuning can enhance a model's ability to produce desired outputs (e.g., feedback for written tasks), thereby reducing the dependence on a user's prompt engineering skills (Jacobsen and Weber, 2023). However, evaluating the performance of LLMs remains a significant challenge (Chang et al., 2024), particularly in the context of AFG, where standard evaluation metrics may not apply, and benchmarks or ground truths are often lacking. van der Lee et al. (2019) emphasize that human evaluation remains the gold standard for assessing the quality of LLM outputs, providing a thorough summary of best practices for such evaluations. Furthermore, consistent with human-centered design principles (Renz and Vladova, 2021), involving expert educators and students in developing and evaluating LLM-based educational tools is crucial. This approach not only ensures that outputs are aligned with the needs of instructors and students but also improves interpretability, thereby mitigating ethical concerns (Yan et al., 2024).

To address existing gaps, this study investigates the potential of fine-tuning a pre-trained LLM for automatically generating feedback messages that align with the characteristics of effective feedback identified in the literature. Specifically, the LLM was fine-tuned on a small set of carefully curated examples to generate feedback on responses to the Casper test, a high-stakes situational judgment test (SJT) designed to assess social intelligence skills. Given the nature of the test, it was crucial for the model to generalize beyond the provided examples, adapting its feedback across different items and responses while accurately inferring character traits implied in the answers. The quality of the generated feedback, particularly its alignment with the structural qualities of effective feedback, was evaluated by two independent judges using a detailed rubric. To gain a comprehensive understanding of model performance, assessment experts also participated in the evaluation process, providing ratings on the quality of feedback content produced by our fine-tuned model. Additionally, a participant study was conducted, where individuals assumed the role of test-takers, interacted with the final fine-tuned model, and then expressed their satisfaction with the received feedback through a survey.

2. Related Work

2.1. Best Practices for Effective Feedback

Feedback provides learners with crucial insights into their performance and understanding as they work toward their goals. Recognized as a powerful catalyst for student learning (Hattie and Timperley, 2007), feedback has even been described as the "cornerstone of all learning" (Stephen Colbran and Colbran, 2017, p. 6). This assertion is supported by extensive research demonstrating that effective feedback can lead to a range of positive student outcomes, including enhanced performance, sustained motivation (Koenka and Anderman, 2019), and the development of effective learning strategies (Matcha et al., 2019).

However, it is important to note that not all feedback is equally effective. Both student preferences and empirical evidence indicate that feedback is most impactful when it is timely, actionable, and personalized (Hattie and Timperley, 2007; Zhang and Hyland, 2018). Feedback that arrives late is often perceived as less relevant, reducing its ability to motivate students and support the achievement of learning goals (Jia et al., 2022). While feedback can address various aspects of performance (Hattie and Timperley, 2007), this study focuses specifically on task-level feedback, which addresses learners' understanding and performance in relation to a particular task.

VanLehn (2006) identifies three key strategies for presenting feedback: (1) binary feedback, which indicates whether a response is right or wrong; (2) error-specific feedback, which highlights where the student's solution deviates from the correct answer; and (3) solution-oriented feedback, which offers hints and strategies for correcting errors. Research suggests that providing only binary feedback–merely informing students whether their solution is correct or not–can lead to confusion and frustration (D'antoni et al., 2015). In contrast, feedback is most effective when it not only identifies mistakes but also offers deeper insights into students' performance, helping them understand their strengths, recognize areas for improvement, and learn how to enhance their future efforts (Sadler, 2010a).

This type of feedback enhances conceptual understanding by not only identifying errors but also helping learners comprehend why their response is incorrect and guiding them toward actionable steps for improvement (D'antoni et al., 2015). Both students and teachers agree that the primary purpose of feedback is to facilitate the improvement of future performance (Dawson et al., 2019). Therefore, effective feedback goes beyond simply informing students about the accuracy of their efforts; it serves a corrective function, offering clear guidance on how to bridge the gap between current performance and the desired goal (Hattie and Timperley, 2007). The ideal characteristics of effective feedback—timeliness, personalization, and actionable insights—should coexist. For example, early engagement with feedback has been linked to better student outcomes when instructors provide personalized weekly emails that include an overview of current performance, links to relevant materials, and specific suggestions for the next steps (Iraj et al., 2021).

Lastly, effective feedback must not only be delivered effectively but also received with attention and openness (Zhang and Hyland, 2018). Feedback is more than just conveying information about learning and performance; it should be an interactive process that engages students in dialogue (Carless, 2016). Alongside cognitive and behavioral engagement, the emotional response to feedback significantly influences how students receive and act on it, affecting their willingness to engage with the feedback (Storch and Wigglesworth,

2010). Therefore, the emotional tone of feedback is crucial. Negative emotions can undermine motivation and self-confidence, making it difficult for students to reflect and improve (Ferguson, 2011). To offer constructive criticism effectively, it is essential to balance the positive and negative aspects of a student's performance while acknowledging their efforts (Hill et al., 2021). Additionally, feedback should focus on the task rather than the individual, and using the second person can help students see the feedback as a subjective perspective, which encourages reflection rather than defensiveness (Prins et al., 2006).

2.2. Automated Feedback Generation (AFG)

Writing effective feedback is often time-consuming for educators and can be especially impractical in large-scale settings. To address this challenge, researchers have been exploring AFG methods for over a decade. Initially developed in the context of computer science and STEM courses, AFG systems have since expanded to other areas, including language education and the arts. These systems typically incorporate expert knowledge through teacher-provided solutions, libraries of correct answers, common errors, and feedback templates. However, only a few AFG tools effectively integrate data-driven techniques with expert insights, and many fail to deliver feedback customized for specific tasks and individual learner characteristics (Deeva et al., 2021).

The literature on AFG encompasses various feedback forms, including graphs and dashboard visualizations, to convey student performance (Cavalcanti et al., 2021). However, this study focuses specifically on written or textual feedback. To generate such feedback, researchers often employ natural language processing (NLP) techniques. For example, Süzen et al. (2020) applied traditional NLP methods in an introductory computer science course by automatically scoring responses to short open-ended questions. Their approach involved text-mining techniques to compare student responses to model answers, using the number of common words to derive scoring rules and assign marks from 0 to 5. K-means clustering was then used to group similar responses into three categories: excellent, mixed, and weak. From each cluster, a prototype answer was selected, allowing teachers to craft a single feedback message for the prototype of each group. New responses could be automatically scored, clustered, and matched with the corresponding feedback. While this method enables timely feedback delivery, it falls short in personalizing feedback, as it does not account for the unique characteristics and needs of individual responses.

Jia et al. (2022) employed advanced NLP techniques to automatically generate feedback on students' project reports using BART, a pretrained language model based on the encoder-decoder transformer architecture (Lewis et al., 2019). The process began with cross-entropy extraction, an unsupervised summarization technique, to condense student reports to a length manageable by BART. The model was then fine-tuned for AFG using these summarized reports paired with corresponding human-written feedback. A manual evaluation across five dimensions—readability, factuality, suggestions, problem identification, and positive tone—revealed that the system was largely unbiased and achieved near-human performance despite being trained on a relatively small dataset (50-100 examples). The model produced fluent feedback that effectively identified problems while maintaining a positive tone. However, 15.2% of the feedback instances were found to be incorrect or ambiguous, and the model lagged behind human experts in offering actionable suggestions.

With the rapid advancements in generative artificial intelligence (AI), educational researchers are increasingly exploring the potential of these tools to automatically generate immediate, personalized, and scalable feedback. Matelsky et al. (2023), for instance, proposed a framework that utilizes pretrained LLMs for AFG on short open-ended questions. In this approach, teachers remain actively involved by defining the questions and evaluation criteria, which are then used to create a prompt. This prompt is stored in a database and paired with student responses before being processed by the LLM for evaluation. Similarly, Steiss et al. (2024) employed prompt engineering with ChatGPT-3.5 to provide feedback on argumentative essays written by students in grades six through twelve, including both proficient English speakers and learners. After experimenting with various prompts, the authors found that the best results were achieved when the model was instructed to act as a secondary school teacher, offering "2-3 pieces of specific and actionable feedback" (Steiss et al., 2024, p. 4) based on the given evaluation criteria and maintaining a positive tone. Although the LLM, without specific fine-tuning, produced feedback that was relatively close to that of expert teachers, human feedback generally outperformed AI-generated feedback.

Jacobsen and Weber (2023) emphasize the critical role of well-crafted prompts in leveraging generative AI for AFG. Their study found that providing the model with detailed instructions and prompting it to think step by step significantly reduced the occurrence of hallucinations—factually incorrect statements—and improved the overall quality of the feedback compared to using less detailed prompts. In their evaluation, AI-generated feedback was compared to that written by human educators with varying levels of expertise. Remarkably, ChatGPT, with version GPT-4, outperformed novice educators and nearly matched the performance of expert teachers, even surpassing them in three of the nine evaluation categories. However, despite using a high-quality prompt, one of the 20 generated feedback instances was notably subpar. This underscores the inherent unpredictability of AI outputs and highlights the ethical considerations that must be addressed when integrating AI into educational practices. Unlike Jacobsen and Weber (2023), Azaiz et al. (2024) found no differences in the quality of model outputs when using different prompts. They obtained more consistent and structured outputs when using GPT-4, compared to its earlier 3.5-turbo counterpart. However, even the latest version of the model generated fully correct and complete feedback on just over half of the instances, with the remaining 48% of feedback containing misclassifications, redundancies, or inaccurate explanations.

2.3. Current Study

This study investigates the potential of fine-tuning both open-source and proprietary pretrained LLMs to generate feedback messages that align with the characteristics of effective feedback identified in the literature. Uniquely, our research focuses on generating feedback for responses to situational judgment questions from the Casper test (Acuity Insights, 2024), which assesses soft skills such as social intelligence and professionalism. Given the subjective nature of these questions—where no single "correct" answer exists, and significant variability is observed between items and responses—the model's ability to generalize beyond the training data is crucial for producing high-quality feedback. Despite task-specific fine-tuning, the model must be capable of handling this variability to be effective. This study is one of the first to apply fine-tuning techniques to LLMs for AFG, contributing to the growing

body of research on generative AI in education. By detailing the fine-tuning process for a specific use case, we hope to provide valuable insights for other researchers and encourage further exploration of AI-driven techniques to develop more robust and effective educational tools.

3. Methodology

3.1. Data Source

The data for this study came from the Casper test–an SJT widely adopted by higher education institutions, particularly in healthcare and education programs, as part of their admissions process. Casper evaluates different aspects of social intelligence and professionalism (e.g., communication, empathy, self-awareness, and resilience). Each item presents one scenario followed by three open-ended questions. After a 30-second reflection period, applicants have up to five minutes to type their responses. Human raters score responses holistically on a scale of 1 to 9, guided by scoring criteria emphasizing expected themes and qualities rather than grammar or writing style. The scenarios can be delivered as text or short videos; however, this study focuses exclusively on text-based responses. Scenarios present applicants with very diverse problems (e.g., navigating workplace conflicts or reflecting on a past challenge) to which no strictly correct answers exist. Consequently, responses vary significantly across items and between applicants answering the same item. Scoring guidelines, therefore, are flexible and emphasize broad thematic evaluation rather than adherence to rigid criteria, making it impractical to create a one-size-fits-all feedback template.

The dataset available to train and evaluate the AFG models consisted of 211,058 written responses to 103 unique text-based scenarios. It contains the soft skills assessed, the scenario text, applicants' responses, two sets of scoring guidelines, and scores assigned by human raters. The first guideline, called the "guiding background," provides detailed context about the focal skills assessed, their relevance to the scenario, and how they should emerge in responses. The second guideline, "guiding questions," distills this background into three to four concise questions, such as "Did the applicant demonstrate [skill]?" or "Did they consider [topic]?" These elements, either in their original or revised form, were employed to fine-tune pre-trained LLMs for AFG through multiple iterations.

3.2. Model Training

Supervised fine-tuning relies on labeled data to teach the foundational LLM to produce outputs that align with user expectations. In this study, no predefined examples of "ideal feedback" were available, so the training dataset had to be built from scratch. While larger training datasets offer more learning opportunities to the model, recent studies indicate that a smaller, high-quality dataset can still yield strong performance (Jha et al., 2023). The optimal number of examples varies depending on the model and its use case, but OpenAI suggests that 50-100 examples are often sufficient to see clear improvements in GPT models (OpenAI, n.d.).

To begin with, we created 100 feedback messages based on responses to 12 different Casper scenarios, ensuring a diverse selection of competencies and performance levels. As

the project progressed, additional examples were added, bringing the final training dataset to 124 examples. The Casper items were randomly chosen to represent a wide range of skills and score distributions, enabling the model to learn how to generate feedback that is both personalized and applicable across varying competencies and performance tiers.

Table 1 provides a detailed breakdown of the examples used for training, categorized by score and scenario. Each feedback message was carefully crafted to align with the key characteristics of effective feedback outlined in the literature. All messages were written in the second person (Prins et al., 2006) to foster a more personal and engaging tone. They balanced positive and negative aspects (Hill et al., 2021) while maintaining a supportive tone, offering actionable suggestions, and integrating evaluation guidelines. Efforts were also made to ensure that the feedback addressed the unique qualities of each response (Hattie and Timperley, 2007), further enhancing its relevance and personalization.

Table 1: Distribution of Training Examples by Scenario and Score

Scenario	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Total**
A	1	2	2	2	1	1	1	0	1	11
B	2	1	2	3	2	2	3	1	1	17
C	1	2	2	1	2	1	1	2	1	13
D	1	1	1	1	1	1	2	1	1	10
E	2	1	1	2	2	1	1	2	1	13
F	1	0	1	1	1	1	1	0	1	7
G	1	1	2	2	3	1	1	1	1	13
H	1	1	0	1	1	1	1	0	1	7
I	0	2	1	0	1	1	1	0	1	7
L	1	0	1	1	1	1	1	1	0	7
M	1	1	0	1	1	1	1	0	1	7
N	1	2	1	2	1	2	1	1	1	12
Total	13	14	14	17	17	14	15	9	11	**124**

We utilized GPT-3.5-Turbo to develop the AFG model. GPT-3.5-Turbo, an enhanced version of GPT-3 and GPT-3.5, is designed to offer a balance between performance and efficiency. With approximately 20 billion parameters (Wodecki, 2023), it presents a cost-effective option, particularly when compared to the pricing of the larger GPT-4 and GPT-4-Turbo models. The specific version used in this study, GPT-3.5-Turbo 1025, has been optimized for greater accuracy in adhering to specified response formats. This model supports outputs up to 4,096 long tokens and is available for fine-tuning via the OpenAI API.

GPT-3.5-Turbo was fine-tuned for AFG using an instruction-tuning approach. After hand-crafting personalized feedback messages, these examples were employed as outputs to construct the instruction-tuning dataset. In this method, the instruction plays a pivotal role in the training data, explicitly defining the task and guiding the model on how to execute it. To optimize performance, the wording of the instruction was adjusted across various iterations of the fine-tuning process. Before submitting the fine-tuning job, the dataset had to be formatted to align with the structure required by the OpenAI API. Each example

was transformed into a chat-style interaction between the user and the system, where the instruction and context were integrated into the user's prompt (see Appendix A for the sample prompts).

3.3. Model Evaluation

LLM outputs can be assessed across three broad dimensions: (1) linguistic quality, (2) information accuracy, and (3) utility (Celikyilmaz et al., 2021). Linguistic quality encompasses factors like grammatical correctness, fluency, and vocabulary use, which are often measured through automated techniques using metrics such as readability scores and lexical diversity. However, many of these metrics either require a reference text or are unsuitable for tasks that allow significant variation in responses (Celikyilmaz et al., 2021). Both challenges are relevant in this context, where responses vary widely, and there is no singular 'correct' feedback for any given response. Consequently, human evaluation remains the gold standard for determining whether LLM outputs achieve the desired qualities (van der Lee et al., 2019).

To evaluate the effectiveness of our AFG model in generating high-quality feedback, we created a detailed scoring rubric and involved two independent judges, along with Casper experts, to assess feedback on responses not seen during training. The rubric covered eight criteria: (1) linguistic quality, (2) factual accuracy, (3) personalization, (4) actionability, (5) affective tone, (6) use of second-person language, (7) adherence to evaluation criteria, and (8) focus/content coverage. The first six criteria, which pertain to the structure of the feedback, were assessed by the independent judges. The last two criteria, which focus on content alignment and comprehensiveness, were evaluated by the Casper experts. This approach ensured a thorough evaluation from both a structural and content-based perspective.

Two judges evaluated 59 feedback messages generated by the AFG model on unseen Casper responses, randomly selected from the dataset to ensure a representative sample of all score ranges. Specifically, six samples were chosen for each score between 1 and 5, and five samples for each score between 6 and 9. The choice to include slightly more low-score responses was driven by the fact that during training, we observed that the fine-tuned model seemed to have more difficulties in these instances, which are also the ones that might be most in need of feedback. After the judges reached close to perfect inter-rater agreement on the independent evaluation of 23 samples, inconsistencies were resolved, and each rater evaluated a unique set of 18 additional generations. Overall, the agreement ranged between 82.6% and 100%, indicating near-perfect agreement between the two judges.

Additionally, to gain a fuller understanding of the model's performance, we conducted a small-scale study where participants interacted with the AFG model as test-takers (see Appendix B), received immediate feedback on their responses, and provided their satisfaction levels through an online survey delivered through Qualtrics. Aligned with the rubric criteria, the survey comprised six sections (three to five questions per section): linguistic quality, factuality, details, personalization, actionability, and affective tone (i.e., balance of strengths and flaws). Within each section, participants used a 6-point Likert scale (1 = completely disagree; 2 = disagree; 3 = slightly disagree; 4 = slightly agree; 5 = agree; 6 = completely agree) to indicate their perceptions. Participants were recruited primarily among undergraduate and graduate students, as this is likely the most represented demographic in the population of Casper applicants. To reach a larger sample size, during the

last two weeks of data collection, the survey was also hosted on Amazon's MTurk, and 91 additional responses were collected. Survey data were processed and analyzed in R (R Core Team, 2023).

4. Results

Our AFG model was first trained using a set of 100 training samples (i.e., feedback examples). A review of the generated feedback indicated that the model produced higher-quality feedback for responses that received middle and high scores but struggled to do the same for low-score responses. Thus, the training data was augmented to include six more examples of feedback on responses that received the lowest possible score (i.e., 1). The model was fine-tuned again using the augmented training dataset without changing the prompt structure. The review of the generated feedback from the updated model suggested that six additional examples were not enough to observe consistent improvement. In the next training iteration, we modified the instruction to use the zero-shot chain of thought (CoT) prompting technique, asking the model to "think step by step". However, this approach also failed to improve the model output adequately. For the next and final iteration, 18 more examples across six scenarios were added to the training set. To aid in creating these samples, zero-shot CoT prompting was leveraged to create a first feedback draft, which was then modified to align them more closely with the desired output. This model cost $1.97 to fine-tune and was retained as our final AFG model. The following sections report the result of the systematic evaluation of the outputs generated by this model.

4.1. Results Based on Structure-Related Criteria

Evaluation of the AFG model's performance on 59 random generations suggests that during fine-tuning, the model picked up the feedback style and structure observed in the training samples (see Figure 1). The model consistently generated feedback pertinent to the context of the scenario and the response with minor linguistic (e.g., misspellings) and factual errors. Most of the evaluated samples were sufficiently personalized. All samples addressed the applicant in the second person, and in almost the totality of instances (84.7%), the feedback pointed out both positive and negative aspects of the response. Similarly, most of the evaluated samples were actionable. However, affective tone and actionability could not be evaluated for 10.1% of the samples because they were generated for responses that received the full score of 9 points and, just like in the received examples, these did not include any suggestions for improvement.

4.2. Results Based on Content-Related Criteria

Our results showed a poor alignment between the independent evaluations of the two Casper experts. However, both experts found that, in the majority of instances (74.6% and 56%), the feedback was aligned or strongly aligned with the evaluation criteria, and, of the 59 samples, only a few generations were flagged as "not at all aligned" with the evaluation criteria (see Figure 2).

Regarding the completeness and focus of the feedback, both experts found more variability in model performance, each finding only a quarter of the generated feedback messages

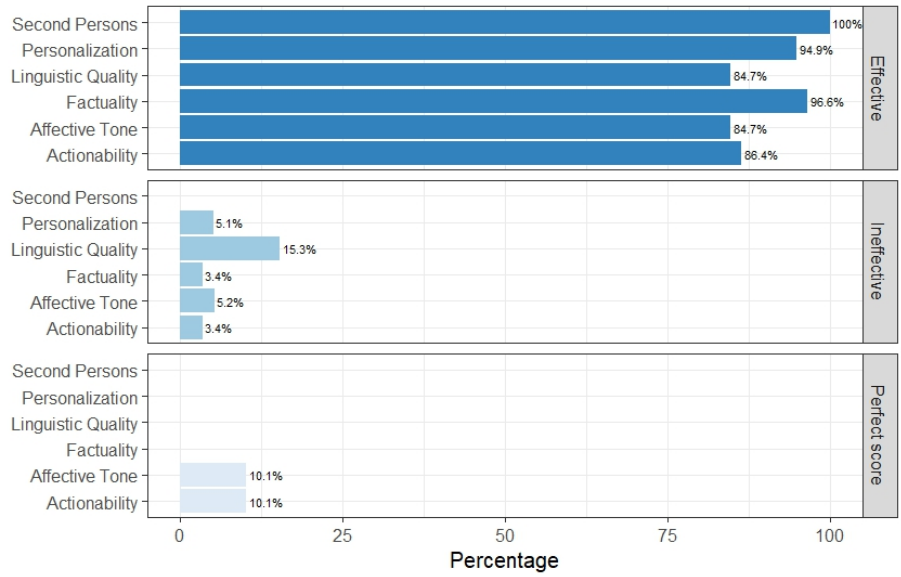

Figure 1: The Proportion of Effective Feedback Messages Based on the Structure-Related Criteria

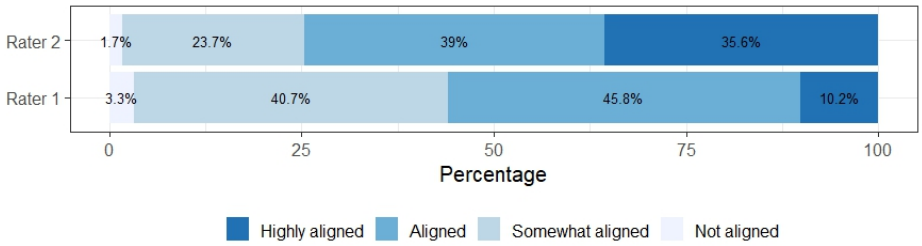

Figure 2: Casper Experts' Evaluation of How Well the Generated Feedback Aligns with Evaluation Guidelines

to be complete and focused (see Figure 3). The frequency of effective and flawed feedback statements did not differ significantly between outputs generated on responses to scenarios seen or unseen during training ($p = .48$).

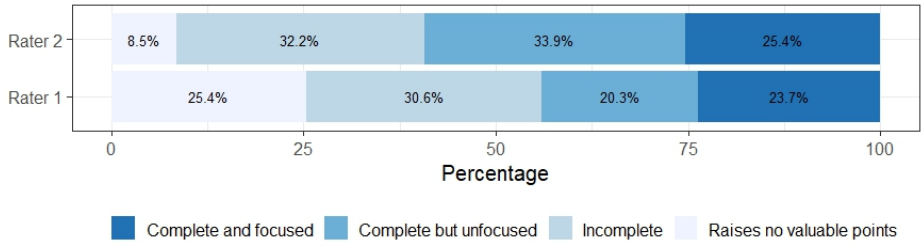

Figure 3: Casper Experts' Evaluation of Completeness and Focus of the Generated Feedback

4.3. User Satisfaction Results

After removing cases with excessive missingness, 164 survey responses were retained. Most respondents (84.8%) expressed satisfaction with the feedback they received. Examining the frequency of responses to individual survey questions, over 70% of participants agreed or strongly agreed that the feedback was clear, grammatically correct, easy to read, and logically structured. Similarly, 70% felt the feedback was relevant to their response. Approximately 60% agreed or strongly agreed that the feedback was personalized to their specific answer. Notably, 78.7% agreed, at least slightly, that the feedback would help them improve their responses. However, in 40.9% of instances, the model failed to meet users' expectations for detailed feedback. Despite this, the model demonstrated a strong ability to provide supportive and balanced responses, with 73.2% agreeing or strongly agreeing that the feedback was encouraging. In less than 20% of cases, participants felt the feedback lacked balance, either being overly negative or positive. Interestingly, 10% of the participants reported that the feedback was not written in the second person. Fisher's exact test found no significant differences in the proportions of satisfied and unsatisfied users, based on whether the test language was their first language ($p = .55$), student status ($p = .35$), or items answered ($p = .71$).

5. Discussion

Effective AFG systems would not only provide valuable support to student learning (Hattie and Timperley, 2007), but also reduce teacher workload and the associated risk of burnout (Jomuad et al., 2021). In this study, we argue that LLMs' ability to generate text and understand context makes them a promising foundation for developing AFG systems that are highly adaptive to different tasks and individual responses. Also, their ability to follow instructions could facilitate the integration of teachers' rules and preferences in the gener-

ated feedback, making AI a bridge between students' need for personalization and support and teachers' involvement in the feedback process.

In this study, we asked the AFG model based on GPT-3.5-Turbo to generate feedback on responses to open-ended questions in a high-stakes SJT measuring social intelligence skills. Similar to Roumeliotis et al. (2024), observations from our study suggest that fine-tuning GPT for AFG leads to better results. Our AFG model was rated quite positively by judges, assessment experts, and survey participants. Although not perfect, our results demonstrate the great potential that LLMs and fine-tuning offer for AFG. This is true, especially when considering the small training size and the fact that feedback messages used to train the model were not written by Casper experts.

The pre-trained GPT adapted highly to the AFG task, picking up the writing style and the effort to create personalized, balanced, and actionable feedback. For example, when the model generated outputs for responses that received a full score of 9, it did not provide any suggestions for improvement; this was also the case in the training data, demonstrating that the model learned that feedback messages on perfect responses were only to point out how the applicant met the evaluation criteria. Moreover, despite the high diversity in items and responses, the model seems successful in generalizing using only a small fine-tuning dataset spanning the full range of assessment constructs and performance levels. However, similar to all other studies using LLMs for AFG, there remained cases where model performance was suboptimal, making our model "useful but fallible" (Matelsky et al., 2023, p. 2). Albeit not hindering the understanding of the message, these linguistic mistakes undermine the validity of the feedback and might also be detrimental to students, for example, in the context of language education or young children who are still developing their writing skills.

Also, while the model generally offers suggestions for improving responses, these recommendations are often not sufficiently comprehensive or focused, as noted by test experts and survey respondents. Some participants wanted more directive feedback, such as specific examples or high-scoring responses. However, this was intentionally avoided, as Casper measures personal character traits, and overly prescriptive feedback could undermine the authenticity of responses and increase the risk of faking. This tension between participants' desire for detailed feedback and the need to preserve test security should be considered in interpreting survey results and in the future development of AFG tools for test preparation.

5.1. Limitations and Future Research

This study has several limitations. First, it did not compare the effectiveness of fine-tuning versus prompting for AFG. Prior research suggests that fine-tuning may enhance only superficial stylistic elements rather than improving logical reasoning. Second, the validity and generalizability of the survey results are limited by the small, non-representative sample, and some participants' unfamiliarity with the high-stakes nature of the Casper test, which may have affected their responses and perception of the feedback. Third, the participants' responses were not scored, depriving the model of an important cue for providing targeted feedback. This may have also contributed to dissatisfaction among some users who felt their feedback lacked actionable suggestions for improvement.

Our findings suggest that improvements in both the content and structure of automated feedback are still needed, including optimizing hyperparameters and enhancing data prepro-

cessing. Future research should involve educators to create larger, more relevant datasets and explore whether the benefits of fine-tuning outweigh the costs compared to prompting out-of-the-box models. Furthermore, we recommend future research expanding AFG systems to other tasks and languages beyond English.

Acknowledgments

The authors want to thank Acuity Insights for sharing the Casper dataset and participating in the evaluation of automatically generated feedback statements in this study. This study was conducted without the support of any third-party funding or sponsorship. The authors did not receive any external financial assistance or resources from organizations, institutions, or individuals in connection with this research. The authors declare that they have no financial interests or financial relationships with any entities that could be perceived as influencing the outcomes of this study. This research was conducted independently, with no financial ties to external parties.

References

Acuity Insights. *Casper Technical Manual*, 2024. URL https://acuityinsights.com/resource/casper-technical-manual/. Accessed: 2024-09-02.

Imen Azaiz, Natalie Kiesler, and Sven 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, July 2024. doi: http://dx.doi.org/10.1145/3649217.3653594.

David Boud and Phillip Dawson. What feedback literate teachers do: an empirically-derived competency framework. *Assessment & Evaluation in Higher Education*, 48(2):158–171, 2023. doi: https://doi.org/10.1080/02602938.2021.1910928.

David Carless. Feedback as dialogue. *Encyclopedia of Educational Theory and Philosophy*, 1:286–289, 2016. doi: https://doi.org/10.1007/978-981-287-532-7_389-1.

Anderson Pinheiro Cavalcanti, Arthur Barbosa, Ruan Carvalho, Fred Freitas, Yi-Shan Tsai, Dragan Gašević, and Rafael Ferreira Mello. Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, 2:100027, 2021. ISSN 2666-920X. doi: https://doi.org/10.1016/j.caeai.2021.100027.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of text generation: A survey, 2021. URL https://arxiv.org/abs/2006.14799.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3), mar 2024. ISSN 2157-6904. doi: https://doi.org/10.1145/3641289.

Loris D'antoni, Dileep Kini, Rajeev Alur, Sumit Gulwani, Mahesh Viswanathan, and Björn Hartmann. How can automatic feedback help students construct automata? *ACM Trans. Comput.-Hum. Interact.*, 22(2), March 2015. ISSN 1073-0516. doi: https://doi.org/10.1145/2723163.

Phillip Dawson, Michael Henderson, Paige Mahoney, Michael Phillips, Tracii Ryan, David Boud, and Elizabeth Molloy. What makes for effective feedback: staff and student perspectives. *Assessment & Evaluation in Higher Education*, 44(1):25–36, 2019. doi: https://doi.org/10.1080/02602938.2018.1467877.

Galina Deeva, Daria Bogdanova, Estefanía Serral, Monique Snoeck, and Jochen De Weerdt. A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education*, 162:104094, 2021. ISSN 0360-1315. doi: https://doi.org/10.1016/j.compedu.2020.104094.

Peter Ferguson. Student perceptions of quality feedback in teacher education. *Assessment & Evaluation in Higher Education*, 36(1):51–62, 2011. doi: https://doi.org/10.1080/02602930903197883.

John Hattie and Helen Timperley. The power of feedback. *Review of Educational Research*, 77(1):81–112, 2007. doi: https://doi.org/10.3102/003465430298487.

Jennifer Hill, Kathy Berlin, Julia Choate, Lisa Cravens-Brown, Lisa McKendrick-Calder, and Susan Smith. Exploring the emotional responses of undergraduate students to assessment feedback: Implications for instructors. *Teaching and Learning Inquiry*, 9(1): 294—-316, 2021. doi: https://doi.org/10.20343/teachlearninqu.9.1.20.

Hamideh Iraj, Anthea Fudge, Huda Khan, Margaret Faulkner, Abelardo Pardo, and Vitomir Kovanović. Narrowing the feedback gap: Examining student engagement with personalized and actionable feedback messages. *Journal of Learning Analytics*, 8(3):101–116, 2021. doi: https://doi.org/10.18608/jla.2021.7184.

L. J. Jacobsen and 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, September 29 2023. Preprint.

Aditi Jha, Sam Havens, Jeremy Dohmann, Alex Trott, and Jacob Portes. Limit: Less is more for instruction tuning across evaluation paradigms, 2023. URL https://arxiv.org/abs/2311.13133.

Qinjin Jia, Mitchell Young, Yunkai Xiao, Jialin Cui, Chengyuan Liu, Parvez Rashid, and Edward Gehringer. Insta-reviewer: A data-driven approach for generating instant feedback on students' project reports. In *Proceedings of the 15th International Conference on Educational Data Mining*, pages 5–16. International Educational Data Mining Society, July 2022. doi: https://doi.org/10.5281/zenodo.6853099.

Perlito D Jomuad, Leah Mabelle M Antiquina, Eusmel U Cericos, Joicelyn A Bacus, Juby H Vallejo, Beverly B Dionio, Jame S Bazar, Joel V Cocolan, and Analyn S

Clarin. Teachers' workload in relation to burnout and work performance. *International journal of educational policy research and review*, 8(2):48–53, 2021. doi: https://doi.org/10.15739/IJEPRR.21.007.

Alison C. Koenka and Eric M. Anderman. Personalized feedback as a strategy for improving motivation and performance among middle school students. *Middle School Journal*, 50 (5):15–22, 2019. doi: https://doi.org/10.1080/00940771.2019.1674768.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019. URL https://arxiv.org/abs/1910.13461.

Wannisa Matcha, Dragan Gašević, Nora'Ayu Ahmad Uzir, Jelena Jovanović, and Abelardo Pardo. Analytics of learning strategies: Associations with academic performance and feedback. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, LAK19, pages 461–470, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362566. doi: https://doi.org/10.1145/3303772.3303787.

Jordan K. Matelsky, Felipe Parodi, Tony Liu, Richard D. Lange, and Konrad P. Kording. A large language model-assisted education tool to provide feedback on open-ended responses, 2023. URL https://arxiv.org/abs/2308.02439.

Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas W Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna 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: 100199, 2024. doi: https://doi.org/10.1016/j.caeai.2023.100199.

Frans J Prins, Dominique MA Sluijsmans, and Paul A Kirschner. Feedback for general practitioners in training: Quality, styles, and preferences. *Advances in Health Sciences Education*, 11:289–303, 2006. doi: https://doi.org/10.1007/s10459-005-3250-z.

George Pu, Anirudh Jain, Jihan Yin, and Russell Kaplan. Empirical analysis of the strengths and weaknesses of peft techniques for llms, 2023. URL https://arxiv.org/abs/2304.14999.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL https://www.R-project.org/.

André Renz and Gergana Vladova. Reinvigorating the discourse on human-centered artificial intelligence in educational technologies. *Technology Innovation Management Review*, 11:5–16, 05/2021 2021. ISSN 1927-0321. doi: http://doi.org/10.22215/timreview/1438.

Konstantinos I. Roumeliotis, Nikolaos D. Tselikas, and Dimitrios K. Nasiopoulos. LLMs in e-commerce: A comparative analysis of GPT and LLaMA models in product review evaluation. *Natural Language Processing Journal*, 6:100056, 2024. doi: https://doi.org/10.1016/j.nlp.2024.100056.

D. Royce Sadler. Beyond feedback: developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, 35(5):535–550, 2010a. doi: https://doi.org/10.1080/02602930903541015.

Royce D Sadler. Beyond feedback: Developing student capability in complex appraisal. *Assessment and Evaluation in Higher Education*, 35(5):535–550, 2010b. doi: https://doi.org/10.4324/9781315872322.

Jacob Steiss, Tamara Tate, Steve Graham, Jazmin Cruz, Michael Hebert, Jiali Wang, Youngsun Moon, Waverly Tseng, Mark Warschauer, and Carol Booth Olson. Comparing the quality of human and chatgpt feedback of students' writing. *Learning and Instruction*, 91:101894, 2024. ISSN 0959-4752. doi: https://doi.org/10.1016/j.learninstruc.2024.101894.

Anthony Gilding Stephen Colbran and Samuel Colbran. Animation and multiple-choice questions as a formative feedback tool for legal education. *The Law Teacher*, 51(3): 249–273, 2017. doi: https://doi.org/10.1080/03069400.2016.1162077.

Neomy Storch and Gillian Wigglesworth. Learners' processing, uptake, and retention of corrective feedback on writing: Case studies. *Studies in Second Language Acquisition*, 32 (2):303–334, 2010. doi: https://doi.org/10.1017/S0272263109990532.

Neslihan Süzen, Alexander N. Gorban, Jeremy Levesley, and Evgeny M. Mirkes. Automatic short answer grading and feedback using text mining methods. *Procedia Computer Science*, 169:726–743, 2020. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2020.02.171. Postproceedings of the 10th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2019 (Tenth Annual Meeting of the BICA Society), held August 15-19, 2019 in Seattle, Washington, USA.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. Best practices for the human evaluation of automatically generated text. In Kees van Deemter, Chenghua Lin, and Hiroya Takamura, editors, *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan, 2019. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/W19-8643.

Kurt VanLehn. The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, 16(3):227–265, 2006. URL https://content.iospress.com/articles/international-journal-of-artificial-intelligence-in-education/jai16-3-02.

Dylan Wiliam. What is assessment for learning? *Studies in Educational Evaluation*, 37(1): 3–14, 2011. doi: https://doi.org/10.1016/j.stueduc.2011.03.001.

B. Wodecki. AI news roundup: Microsoft may have leaked ChatGPT parameters. *AI Business*, November 3 2023. URL https://aibusiness.com/verticals/-ai-news-roundup-microsoft-may-have-leaked-chatgpt-parameters.

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. Practical and ethical challenges of

large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112, 2024. doi: https://doi.org/10.1111/bjet.13370.

Zhe (Victor) Zhang and Ken Hyland. Student engagement with teacher and automated feedback on l2 writing. *Assessing Writing*, 36:90–102, 2018. ISSN 1075-2935. doi: https://doi.org/10.1016/j.asw.2018.02.004.

Appendix A

The following shows the initial prompt structure that we used to fine-tune GPT-3.5-Turbo.

```
{"messages": [
    {"role": "system", "content":
     "You are a tutor that generates feedback on
     responses to situational judgment items based on
     provided criteria."},

    {"role": "user", "content":
     "Instruction: Generate feedback for the answer to the following questions:
      [questions] + Base your feedback on the following criteria:
      [guiding summary] + Disregard spelling, grammar, and style +
      [response] + This answer got a score of [score] out of 9."},

    {"role": "assistant", "content": "[feedback]"}
]}
```

The following shows the final prompt structure that we used to fine-tune GPT-3.5-Turbo.

```
{"messages": [
    {"role": "system", "content": "You are a tutor that generates feedback on
     responses to situational judgment items based on provided criteria."},

    {"role": "user", "content": "Instruction: Generate feedback for the answer to
    the following questions: [questions] + Base your feedback on the following
    criteria: [guiding summary] + Disregard spelling, grammar, and style. +
    When you generate feedback, let's think step by step. Explain your reasoning
    process and how your feedback relates to the answer and the evaluation criteria.
    + [response] + This answer got a score of [score] out of 9."},

    {"role": "assistant", "content": "[feedback]"}
]}
```

Appendix B

Figure 4 shows a screenshot of one of the Casper tasks shared with participants.

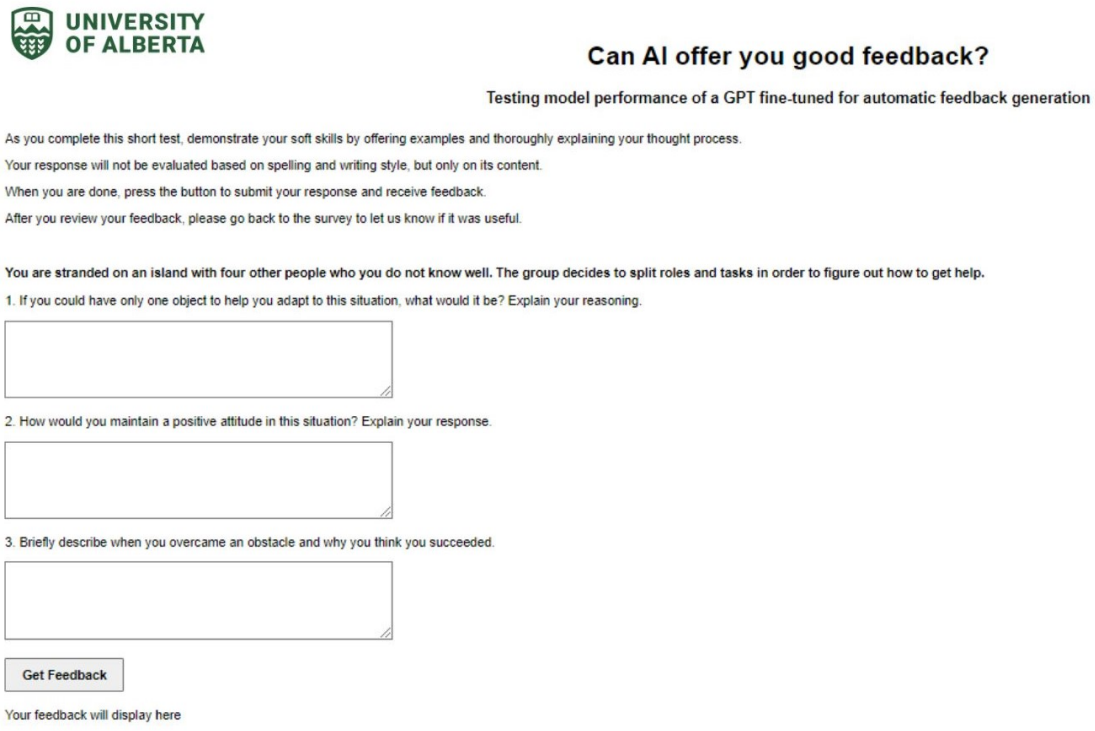

Figure 4: A Screenshot of a Casper Task Shared with Participants