**Can AI Offer Effective Feedback for Open-Ended Questions?**

**Results from Fine-Tuned LLMs for Automatic Feedback Generation**

by

Elisabetta Mazzullo

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Education

in

Measurement, Evaluation, and Data Science

Department of Educational Psychology
University of Alberta

## Abstract

The value of timely, personalized, detailed, and actionable feedback for learning is widely recognized; yet its provision is time-consuming and unfeasible in large-scale contexts. Current automatic feedback generation (AFG) methods lack customization and educator involvement. Large language models (LLMs), with their abilities to analyze and generate text, could address these limitations: existing studies using out-of-the-box LLMs and prompting strategies achieved positive results, but space for improvement remains. Parameter efficient fine-tuning can customize pre-trained LLMs using limited data and memory. Also, open-source LLMs can offer cost and privacy advantages with comparable performance to proprietary models. This study explores fine-tuning both closed and open-source LLMs for AFG. Using Meta's Llama-2-7B and OpenAI's GPT-3.5-turbo, we generated feedback for open-ended responses to situational judgment questions. The models were fine-tuned on a small set of hand-crafted feedback examples, using prompting strategies in the training instruction. The final model was evaluated by two independent judges and test experts. In addition, a user satisfaction survey was conducted for participants to interact with the model as test takers and evaluate their feedback. Our findings suggest that fine-tuning produces better results when using GPT, and it underscores the importance of effective training instructions. The fine-tuned model achieved a high user satisfaction (84.8%) and largely met desired structural qualities (72.9%). Also, it successfully generalized across different items, and provided feedback aligned with the given instructions, functioning similarly well, regardless of performance level, English learner status, or whether the respondent is currently a student. However, there remain instances where outputs contain linguistic mistakes, fail to provide focused suggestions, or feel quite generic. Suggestions on how these issues might be addressed, implications, and ethical considerations are discussed.

## Preface

This thesis is an original work by Elisabetta Mazzullo. No part of this thesis has been previously published. The research project conducted as part of this thesis received research ethics approval from the University of Alberta Research Ethics Board 2, Project Name "Can AI offer you good feedback? - Testing model performance of a GPT fine-tuned for automatic feedback generation" No. Pro00140363, May 24, 2023. The generative artificial intelligence applications or Large Language Models ChatGPT-3.5-turbo and Llama-2 were used as the foundational models for fine-tuning in our research work. Furthermore, one of our fine-tuned GPTs was used to aid in the generation of additional training data.

## Table of Contents

## List of Tables

## List of Figures

# Chapter 1: Introduction

## Background

The Only Thing I Know About Scientists.

*A scientist asked me, who are you?*

*I told her, I'm a dog in front of my master.*

*She smiled, then tossed a stick for me to catch.*

*And I fetched it.*

- Code-Davinci (2023)

With these words, one of the first publicly-accessible large language models (LLMs) poetically describes itself. Recent technological innovations have largely improved the generative abilities of Artificial Intelligence (AI) models, expanding the potential applications for these tools. With the increasing ease of access to LLMs, researchers in every field are playing master: throwing sticks in every direction to explore what the dog can do, and looking for ways to teach it new tricks. In the context of education, LLMs are being tasked to automatically generate items (Tan et al., 2024), score students' responses (Latif & Zahi, 2024), offer them feedback (Meyer et al., 2024; Phung et al., 2024; Steiss et al., 2023) and tutoring (Scarlatos, 2024). Through these applications, LLMs could reduce instructors' workload while integrating their knowledge and preferences into their outputs and offer students highly personalized learning experiences and interpretable insights about their performance (Mazzullo et al., 2023).

Decades of research testify to the importance of feedback in education (Hattie & Timperley, 2007) and advocate for a switch from assessment of learning to assessment for learning (Wiliam, 2011). The literature encourages educators to provide students with feedback that is timely, personalized, and detailed (Hattie & Timperley, 2007), that actively involves them

in dialogue (Carless, 2016), and provides suggestions for improvement (Sadler, 2010). However, writing feedback that meets these standards poses high demands on teachers (Boud & Dawson, 2021), who already face heavy workloads and the associated risk of burnout (Jomuad et al., 2021). In addition, and of concern for the present study, the delivery of such feedback is completely unfeasible in the context of large-scale assessment and large classrooms, such as the increasingly popular massive online courses. This s Feedback remains highly desirable in these contexts as well, especially if the assessment is highly consequential to test takers, or if online learning comes at the price of limited or no access to instructors' or peers' support.

With their powerful language understanding and generative abilities, LLMs might offer a solution to deliver timely and personalized feedback at scale. In addition, given their instruction-following capabilities, LLMs allow educators to maintain a level of control in the feedback process, as they can provide directions or examples to lead the model toward the desired output (e.g. Meyer et al., 2024). Existing research on the application of LLMs for automatic feedback generation (AFG) reports promising results (Matelsky et al., 2023); nevertheless, the field is new and changing rapidly, and many areas remain unexplored.

**Gaps in the Literature**

Most studies so far have relied on the massive proprietary LLMs such as OpenAI's GPT models; however, smaller open-source LLMs (e.g., Llama, Mistral) are also available, and they might represent a more accessible solution for researchers and small organizations while still achieving high performance (Bergmann, 2023). In addition, these studies often used GPT in its publicly available ready-to-use chat version (i.e., ChatGPT), only leveraging prompting techniques to shape the feedback output. Recently introduced fine-tuning techniques offer a cost-effective way to tailor pre-trained LLMs for a specific task (Pu et al., 2023), without necessarily

requiring a large tuning dataset (Jha et al., 2023). Fine-tuning might improve the model's ability to generate desired outputs and reduce reliance on the instructor's prompt engineering abilities (Jacobsen and Weber, 2023). Lastly, the evaluation of LLM performance remains challenging (Chang et al., 2024), especially in AFG, where many automatic evaluation metrics cannot be applied, and no benchmark and sometimes no ground truth are available. Van Der Lee et al. (2019) noted that human evaluation remains the gold standard for judging the quality of LLM outputs, and they offer a comprehensive summary of good practices for conducting such evaluations. In addition, in line with principles for human-centered design (Renz & Vladova, 2021), it is important to involve both expert educators and students in the evaluation of LLM-based educational tools. This would not only increase the alignment of outputs with instructors' and students' needs but also increase interpretability, thereby reducing ethical concerns (Yan et al., 2023).

To address these gaps, the present study explores the possibility of fine-tuning both open and closed-source pre-trained LLMs for the automatic generation of feedback messages that meet the characteristics of effective feedback identified in the literature. Specifically, the LLMs were fine-tuned on a small set of hand-crafted examples to generate feedback on responses to the Casper test, a high-stakes situational judgment test (SJT) assessing social intelligence skills. The purpose of this feedback is to help applicants improve their Casper responses and thereby their test performance. Given the nature of the test, the model's ability to generalize beyond examples, and across items and answers remains crucial to generate high-quality feedback, together with its ability to infer character traits implicit in the responses. The model's ability to generate outputs aligned with desired structural qualities of effective feedback was evaluated by two independent judges using a rubric. To obtain a thorough understanding of model performance, test experts

involved in the design of Casper also participated in the evaluation process to rate the quality of the feedback generated by our fine-tuned model. In addition, we conducted a study that let participants assume the role of test takers and interact with our final fine-tuned model before expressing their satisfaction with their feedback through a survey.

**Structure of the Thesis**

The following chapter will present the fundamental concepts our study builds on. First, it summarizes literature recommendations for effective feedback, identifying the standards our AFG model should strive to meet. Then, foundational architectures and essential techniques for natural language generation are briefly introduced. Later, we present the state of research on the use of LLMs for AFG, with attention to the methods used and results achieved so far in the field. Chapter 2 offers a detailed overview of the materials and methods used in the present study, from the development of the training dataset to our final AFG model. We present the scope and structure of the Casper test, the pre-trained LLMs we adopted for fine-tuning, and the approaches used to do so. Lastly, we outline the procedures used to evaluate model performance, both through a rubric and a survey study. Observations from the fine-tuning process and evaluation results are reported in Chapter 3. Later, we reflect on the meaning of our findings in relation to the literature, identifying how our study advances knowledge in the field and questions that remain open. Lastly, we take a critical look at our work, as we recognize limitations and directions for future work.

## Chapter 2: Literature Review

### The Importance of and Best Practices for Educational Feedback

Feedback provides learners with information about their performance or level of understanding as they work to reach a goal (Wiggins, 2011). It has long been recognized as a powerful facilitator of student learning (Hattie & Timperley, 2007), so much so as to be described as the "cornerstone of all learning" (Colbran et al., 2016, p.6). This statement is supported by decades of research showing that feedback can support an array of positive outcomes for students, including improved performance, sustained motivation (Koenka & Anderman, 2019), and the development of learning strategies (Matcha et al., 2019).

It should be noted that not all feedback is created equal. Student preferences, backed up by empirical evidence, maintain that feedback is most effective when it is timely, actionable, and personalized (Hattie & Timperley, 2007; Sadler, 2010; Zhang & Hyland, 2018). Tardy feedback is perceived as less relevant, and it is less likely to motivate students to stay on task and to encourage the achievement of learning goals (Jia et al., 2022). While feedback can comment on multiple aspects of performance (Hattie & Timperley, 2007), the present study is specifically concerned with task-level feedback, which focuses on the relationship between learners and a specific task, regarding the understanding and performance on the given task. Vanlehn (2006) identifies three strategies to present this type of information: (1) binary feedback (i.e., right or wrong), (2) error-specific feedback, highlighting where the provided solution diverges from the correct response, and (3) solution-oriented feedback: offering hints and strategies to repair the error in a student's response. Only telling students whether their effort resulted in a successful solution or not is found to be confusing and frustrating for students (D'Antoni et al., 2015). Instead, feedback is most effective when it provides learners with insights into their

performance, helping them identify not only their mistakes, but also their strengths, weaknesses, and how they can do better (Sadler, 2010).

This type of feedback aids conceptual understanding because it does not only point out what is wrong, but it helps learners understand how their response is wrong and identify actionable next steps to improve (D'Antoni et al., 2015). In fact, students and teachers agree that the purpose of feedback is the improvement of students' future performance (Dawson et al., 2018). Therefore, good feedback does not only inform students about the correctness of their efforts, but it assumes a more correctional purpose, giving directions to fill the gap between the current performance and understanding and a desired or expected goal (Hattie & Timperley, 2007). All the ideal characteristics of effective feedback can and should coexist; for example, early engagement with feedback was associated with better student outcomes when instructors sent personalized weekly emails that offered an overview of current performance, along with links to relevant materials and suggestions on what to do next (Iraj et al., 2021).

Lastly, feedback must not only be delivered effectively, but be received attentively (Zhank & Hyland, 2018). In fact, effective feedback does not consist in the mere provision of information about learning and performance, but it is a process where students should actively be engaged in dialogue (Carless, 2016). Together with the cognitive and behavioral components of engagement, the emotions elicited by the feedback impact how students respond to it, influencing acceptance and willingness to work on the feedback (Storch & Wigglesworth, 2010). This makes the affective tone of the message another important aspect to attend to when crafting high-quality feedback. If feedback elicits negative emotions, these can reduce motivation and self-confidence, making it challenging to reflect and act upon it (Ferguson, 2011). Therefore, to deliver constructive criticism, it is advised to balance the positive and negative aspects of a

student's performance and to acknowledge their efforts (Hill et al., 2021). Moreover, feedback should focus on the task product rather than the individual (Hill et al., 2021), and addressing learners in the second person helps students perceive the feedback as a subjective viewpoint rather than an indisputable fact (Prins et al., 2006).

**Automatic Feedback Generation**

Writing feedback messages that meet these characteristics is highly time-consuming for educators (Boud & Dawson, 2021), or entirely unfeasible in large-scale contexts; therefore, for over a decade, researchers have explored methods for automatic feedback generation (AFG). These systems have primarily been developed in the context of computing science and STEM courses, but examples of AFG research extend to other skills and subjects, such as language education and the arts. Expert knowledge is typically integrated in automatic feedback systems in the form of teacher solutions and libraries of correct answers, sets of typical errors, or feedback templates/rules. However, only a minority of these tools incorporate data-driven techniques with expert knowledge, and existing AFG systems often fall short in the provision of feedback personalized on tasks and individual characteristics (Deeva et al., 2021).

The literature on AFG broadly recognizes as feedback a variety of pieces of information that can be presented to students to convey their performance, including graphs, dashboard visualization, and ontologies (Cavalcanti et al., 2021). However, from now on, the present study will uniquely be concerned with feedback in the form of natural language messages. Literature reviews on approaches and methods for AFG for programming exercises (Keuning et al., 2018) and AFG in online learning environments (Cavalcanti et al., 2021) identify four important directions for research: (1) creating feedback that offers correctional hint and actionable suggestions, (2) developing tools focused on instructors, (3) using natural language generation

techniques, (4) seeking students' and instructors' inputs to evaluate the quality of the generated feedback.

**Natural Language Processing and LLMs**

Natural language processing (NLP) is a rapidly evolving branch of AI that combines computational linguistics with statistical methods, machine learning, and deep learning to process and analyze textual data and perform a variety of tasks involving natural language. These tasks include, for example, sentiment analysis, machine translation, and text generation. A fundamental step of many NLP methods is tokenization, where the input text is broken down into smaller units, typically words or sub-words, called tokens. Each token corresponds to a unique integer, so that the raw text data can be converted into a numerical sequence that can be fed into a learning model (Holdsworth, 2024).

Huge advancements in the field were brought about by the introduction of the transformer architecture and of the self-attention mechanism (Vaswani et al., 2017). Transformers are essentially deep learning neural networks (NNs). Based on the different roles that they play in the model's overall function, they can be distinguished into three architectures: encoders, decoders, and encoder-decoders. Encoders aim to understand the input sequence. They focus on processing the input and capturing its meaning and context. To do so, the sequence of integers obtained from the tokenization is converted into dense continuous-valued vectors called input embeddings. These word embeddings capture semantic relationships, modeling relationships between words in a lower-dimensional space. The transformer creates embeddings using a series of layers, each built in the same way. In particular, positional encoding encodes the position of each element in the input sequence as a set of numbers, registering information about the relative position of tokens. In addition, Transformer blocks include a self-attention layer. The

self-attention mechanism allows the model to remember multiple words in a sequence and to create a context-aware representation of the input: the model weighs the contextual importance of different words and phrases in a text, improving its ability to extract relationships between different elements in the text sequence and to handle long-range dependencies. Once the input embeddings have been created, a feed forward NN extracts patterns from the data to make predictions, for example for tasks of sentiment analysis or name entity recognition.

The information extracted by the encoder can also be passed to a decoder. Decoder models generate an output sequence based on the information learned by the encoder, making them well versed for tasks of text generation via next-word prediction. Similar to encoders, they consist of different layers, but, due to their function, the self-attention layer is masked to only attend to the words that have come before the current one, not to those that come after. In addition, decoders sometimes include a cross-attention layer, through which the model continuously checks the input (output of the encoder) as it generates its predictions. Once again, a feed-forward NN learns from these data and makes predictions. Lastly, encoder-decoders combine both encoder and decoder components into a single model. They are used in tasks where the input and output sequences have different lengths or meanings. The encoder understands the input sequence, and the decoder generates the corresponding output sequence. This architecture is often used for translation, text generation, and question-answering tasks. Unlike recurrent NN, which previously dominated NLP techniques, Transformers do not process the input sequentially but can handle information in parallel. This makes them more efficient and particularly good at extracting relationships between words in a sentence or longer text sequence (Vaswani et al., 2017).

Transformers form the architecture of many state-of-the-art LLMs. LLMs are deep learning algorithms trained on a vast corpus of textual data, and they are called large because they include billions or even trillions of parameters. The Generative Pretrained Transformer (GPT) is a language model developed by OpenAI that uses deep learning techniques to generate human-like text in response to a text-based input prompt. Its architecture is based on the decoder-only transformer and on the self-attention mechanisms (Yenduri et al., 2024). The transformer is pre-trained on a vast amount of raw text data using unsupervised learning techniques. During pre-training, thanks to the self-attention mechanism, the model learns the relationships between words and their meaning. It learns to generate utterances auto-regressively via next-word prediction, by predicting and appending subsequent tokens one by one based on the context of the ones that have come before, until a <stop> token is selected. The prompt provides the initial context for the model to begin generating text. Pre-trained LLMs are highly versatile, and capable of performing well across a range of tasks right out of the box. Furthermore, they can be optimized for specific purposes using prompt engineering and fine-tuning techniques (see Chapter 3). Prompt engineering steers the model towards the generation of a desired output without performing any additional training while fine-tuning updates model parameters for the downstream task using task-specific training data.

Building onto the strength of transformers, LLMs seem to offer great potential for educational applications, including the automatic generation of feedback that is tailored to each student response and built around instructors' needs (Kasneci et al., 2023). With their generative and context-understanding abilities, LLMs allow to move past predefined templates (e.g., Ramo-Sotos et al., 2016; Suzen et al., 2020) and hardcore heuristics (e.g., Pardo et al., 2018) for the

generation of feedback, potentially allowing for greater personalization while keeping instructors involved in the process.

**NLP and LLMs for AGF**

An example of the application of traditional NLP for AFG can be found in Suzen et al. (2020). Using text mining techniques, responses to a set of short open-ended questions in an introductory computer science course were automatically scored. Each response was compared to a model answer, and similarity based on the number of common words was computed to derive scoring rules and automatically assign a mark between 0 and 5. Later, K-mean clustering was used to group similar responses into three groups (excellent, mixed, and weak). The authors suggested that from each cluster, a prototype answer be chosen so that teachers can develop one feedback for the prototype answer of each cluster. New answers can then be automatically scored and clustered, and receive the feedback associated with their given cluster. While this approach involves teachers in the process of delivering timely feedback, it does not take into consideration the unique characteristics and needs of each answer, thus failing in terms of personalization of feedback.

Later, Jia et al. (2022) used innovative NLP methods to automatically generate feedback on students' project reports with BART, a pretrained language model based on the encoder-decoder transformer architecture (Lewis et al., 2019). First, the unsupervised summarization technique of cross-entropy extraction shortened student reports to a length accepted by BART; later, the model was trained for AFG using the summarized reports and the associated human-written feedback. A manual evaluation conducted on five dimensions (i.e., readability, factuality, suggestions, problems, and positive tone) concluded that the system was free of bias and that it achieved near-human performance despite the small set of examples used for training (50 and

100). The model was particularly strong in generating fluent feedback that pointed out problems while maintaining a positive tone. However, 15.2% of the evaluated instances were found to be incorrect or ambiguous, and the model was not quite as good as experts in offering suggestions.

Following the rapid improvements in generative AI introduced by the GPT, educational researchers are increasingly exploring how these tools can be used for the automatic generation of immediate, personalized, and scalable feedback. For example, Matelsky et al. (2023) proposed a framework to leverage any pre-trained LLM for the automatic generation of feedback on responses to short open-ended questions. Teachers are kept "in-the-loop" as they define the questions and evaluation criteria. These are used to create a prompt, which is stored in a database and paired with students' responses before they are sent to the LLM for evaluation.

Steiss et al. (2023) used prompt engineering and ChatGPT-3.5 to offer feedback for argumentative essays written by students in grades six to twelve, both proficient and learners in the English language. After trying our different prompts, the authors concluded that the best results were achieved when the model was asked to act as a secondary school teacher, to "provide 2-3 pieces of specific and actionable feedback" (p. 4) based on given evaluation criteria, and to maintain a positive tone. Hence, the model received no additional training on the study task, and it was not offered any examples of the desired response. The only modification made to ChatGPT was in the temperature hyperparameter, which was set very low to restrict the randomness and creativity of the outputs. 198 feedback messages generated with this method were evaluated using a rubric and compared with expert-generated feedback. Overall, the LLM, without any specific training, was able to generate feedback relatively close to that written by expert teachers. Human feedback outperformed feedback generated from ChatGPT in four of the five evaluated aspects: clarity of directions for improvement, accuracy, prioritization of essential

features, and use of a supportive tone. However, the authors argued that these differences are minimal when considering the time savings.

Using a similar prompting method and the later GPT versions, ChatGPT-3.5-turbo, Meyer et al (2024) generated feedback on argumentative essays for English language learning. To guide the model in the generation of high-quality feedback, the authors crafted a very detailed prompt. This specified the grade and language learner status of the students, it asked to evaluate three specific aspects of the response and to provide three hints and three examples for improvement for each of these three aspects. It explained in detail what these should look like and how they should be displayed on a table. In addition, using the one-shot prompting paradigm, the prompt includes one example to further help the model understand what the desired output should look like. To evaluate the effectiveness of this method, grade ten students were asked to write short essays and to revise them either without or after having received the LLM-generated feedback. Essays were automatically scored before and after revisions, and relative improvements in performance were compared between the control and experimental groups. The randomized control trial concluded that the automatic feedback was effective in increasing revision performance, task motivation, and positive emotions. However, the relative improvement in performance was small, and average usefulness as perceived by students hovered around the middle of the scale, encouraging researchers to strive for better results.

Jacobsen and Weber's (2023) study on the use of generative AI for AFG highlights the importance of using well-crafted prompts to obtain high-quality results. They found that including specific instructions and asking the model to "think step by step" reduced hallucinations (i.e., factually incorrect statements) and significantly improved model outputs, compared to the use of two weaker prompts. The automatically generated feedback was

compared to that written by humans with varying levels of expertise. ChatGPT-4 surpassed novices in the task, and performed almost as well as expert educators, outperforming them on three out of the nine categories considered for evaluation. However, even when using the high-quality prompt, one of the 20 generated feedback statements was quite poor, a reminder of the unpredictable shortcomings of AI and of the ethical challenges that come with its application in educational settings.

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 ChatGPT-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, inconsistencies, or inaccurate explanations.

Overall, these tools appear to be "useful but fallible" (Matelsky et al., 2023, p. 2). LLMs seem to achieve mostly positive and encouraging results: even without any task-specific training, they can generate coherent and mostly correct feedback, and they keep educators "in the loop" by allowing them to define evaluation criteria and provide direction to shape the model output. Therefore, these systems could alleviate teachers' workload and represent a valuable resource for student learning. However, they sometimes make inaccurate or incorrect statements, which prevents the educational community from fully trusting them on their own and raises ethical issues about the risk associated with these tools (Jacobsen & Weber, 2023; Yan et al, 2023). To alleviate concerns about misinformation, Phung et al. (2024) leveraged OpenAI's GPTs not only for feedback generation but also as a quality assurance layer. First, GPT-4 was used to generate hints to fix buggy Python codes in programming education. To guide the model towards the

generation of high-quality outputs, the prompt specified that the hint should be one-sentence long, neither too specific nor too abstract, and that it should stimulate thinking to reach the correct solution. In addition, to support the model's reasoning abilities, the prompt also requested an explanation of the bugs. Later, the weaker GPT-3.5 was used to simulate a human student and determine the utility of the automatically generated hint. Results find that including symbolic information (i.e., failing test cases and fixed programs) in the prompt leads to better performance, and human evaluation deemed 95% of the hints generated with this system comparable to those written by tutors.

Other important considerations about the ethical implications of using AI in education, can be summarized into three essential topics: data privacy, equality (i.e., accessibility to stakeholders with different backgrounds), and beneficence/potential harms (e.g., misinformation) (Ferguson et al., 2016). These concerns are exacerbated by the fact that LLMs are "the ultimate black box AI method" (Yenduri et al., 2024, p.6), due to the lack of transparency in the data source, their high model complexity, and their subsequent low interpretability and scarce prediction explicability. Yan et al. (2023) pointed out a lack of consideration for these concerns in studies leveraging LLMs for educational applications, and they advocate for the adoption of a human-centered approach in the process of development and evaluation to make these systems transparent to educators and parents.

**Current Study: Addressing Open Questions in an Evolving Field**

The literature on the use of LLMs for AFG is growing, but still limited. Most studies so far have made use of the massive GPT Chat models owned by OpenAI (Meyer et al., 2024; Phung et al., 2024; Steiss et al., 2023). However, publicly available open-source LLMs (e.g., Llama, Mistral), using fewer parameters, might be able to achieve comparable performance

while offering affordable and manageable solutions for users with limited budgets and infrastructures (Bergmann, 2023). To the best of our knowledge, the use of any of these models for AFG currently remains unexplored. In addition, the majority of studies have generated feedback using pre-existing Chat models and prompting strategies, sometimes altering the temperature hyperparameter (Steiss et al., 2023), but never performing any additional training to adapt the model's weights for a specific task of interest. While this represents an additional cost, some recently introduced methods allow for highly efficient fine-tuning of pre-trained LLMs (Pu et al., 2023). The additional training might lead to improved task performance (Roumeliotis et al., 2024), and it does not always require an extensive training dataset (Jha et al., 2023).

These techniques can be applied to specialize LLMs on educational tasks; for example, Latif and Zahi (2024) found that a fine-tuned GPT-3.5-turbo achieved high accuracy in the automatic scoring of written responses to a science education assessment, significantly outperforming Google's BERT. Moreover, Jacobsen and Weber (2023) warn that the need to design high-quality prompts in order to truly rip the benefits of the generative power of AI models like ChatGPT for AFG might represent a barrier to entry for teachers. Fine-tuning pre-trained LLMs for AFG might alleviate this issue: the specialized model would be inherently predisposed to generate feedback with desired characteristics in the context of a specific task or subject and would tend to produce more consistent outputs, possibly making it less susceptible to wording effects in the prompt (Bergmann, 2024). Lastly, in a landscape where evaluation of the outputs generated by LLMs remains challenging, most studies rely on rubrics and human raters to evaluate model performance. When conducting these evaluations, scholars remind us that it is important to involve expert teachers as well as students in the process (Cavalcanti et al., 2021; Renz & Vladova, 2021; van Der Lee et al., 2019).

To address these gaps and challenges, the present study explores the possibility of fine-tuning both open and closed-source pre-trained LLMs for the automatic generation of feedback messages that meet the characteristics of effective feedback identified in the literature. Peculiar to our research, the LLM is tasked to generate feedback on responses to situational judgment questions from the Casper test (Acuity Insights, 2024). These questions assess soft skills related to social intelligence and professionalism; given their nature, no single right or wrong answer exists, and extensive variability is found both between items and between responses to the same item. Therefore, even though the model will receive task-specific training, its ability to generalize beyond the training data remains crucial for the generation of high-quality outputs. To evaluate model performance, two independent judges checked the structure of the generated feedback against the characteristics of effective feedback using a rubric. In addition, to include a broader range of perspectives, expert Casper evaluators were involved in the process of assessing the content of the generated feedback; a survey study allowed lay users to interact with the model as test takers and express their satisfaction with their own feedback.

Therefore, the present study aims to explore the effectiveness of fine-tuning both open and closed-source LLMs for AFG on a small set of hand-crafted examples. Effective feedback is defined based on evidence from the literature, and we employ two different approaches for the evaluation of model performance, involving multiple subjects of the feedback process. For the first time, the study applies fine-tuning for AFG, adding to the growing literature on the application of AI for educational purposes. We report in detail the process of fine-tuning LLMs for a specific use case, in the hope that other researchers can learn from our experience and to encourage the investigation of other techniques and models for the development of stronger and more effective tools.

## Chapter 3: Methods

After introducing the core concepts of fine-tuning LLMs, the following section presents the dataset and methods used in the present study to generate and evaluate our AFG model. Research with LLMs is often a process of trial and error, and this study was no different. Here we outline the steps that led to the final feedback model, explaining the characteristics of the LLMs that were employed, the rationale for the adoption of specific training techniques, and for changes in the training process made from one iteration to the next. Lastly, the procedures used for evaluation of the feedback generated by our fine-tuned LLM are reported. While recounting all the turns taken along the way will not make for a concise presentation of the methods employed in the final model, sharing the challenges and lessons learned in the process is necessary to give stronger grounding to our final model, and it could better aid other researchers as they undergo their own explorations with fine-tuning LLMs.

### Fine-Tuning LLMs for Feedback Generation

The strength of pretrained LLMs lies in their auto-regressive nature, meaning that foundational models are well-versed in the prediction of the next word in a given sentence until the sequence is complete. This makes the pre-trained model capable of generating linguistically coherent text, but not necessarily able to satisfy specific user needs and requests. To bridge the gap between the model's ability to process and predict language and its ability to perform specific tasks, fine-tuning can be used to adapt a LLM to a specific downstream task. For example, LLMs can be specialized for dialogue, information extraction, classification, sentiment analysis, writing, or arithmetic (Zhang et al., 2023). This is typically done by adapting the parameters and representations of an existing pre-trained LLM, thus leveraging the broad knowledge and stability of a model trained on an extensive dataset as the starting point for a

model tailored to a specific use case. This significantly reduces the computational costs compared to training an LLM from scratch, and it requires a smaller amount of labeled data (Zhang et al., 2023). Through fine-tuning, the model gains new knowledge and generally achieves better results on the downstream tasks it was trained to perform (Ding et al., 2023).

Fine-tuning is generally carried out through a supervised learning approach, although semi- and self-supervised learning techniques also exist. Supervised fine-tuning (SFT) requires the use of a dataset that provides the model with specific knowledge it might lack or with examples of the task to be learned. In a typical fine-tuning process, this dataset consists of numerous input–output pairs for a specific task. For example, if a pre-trained LLM was to be tuned for poetic translations from Italian to English, the example dataset would contain input sentences in the original language, and output translations in the target language (e.g., input: “e quindi uscimmo a riveder le stelle” output: “and thence we came forth to see again the stars”). Thanks to these labeled data, new learning happens as model parameters are updated, and this can be achieved through different approaches.

Full fine-tuning (FFT) updates all weights of the foundational LLM, including its lower-level representation. This is typically done when the downstream task requires substantial adaptation as it is very different from the data used for pre-training. To avoid destabilizing changes, model hyperparameters that influence the learning process can be adjusted based on their pre-training specifications (e.g., setting a small learning rate to avoid dramatic forgetting and overfitting). However, FFT is extremely costly, especially as LLMs carry an increasingly large number of parameters. In addition, when the pre-training data and the task-specific data share similarities, the lower-level representation of the foundational LLM is still relevant, making the effort to update those parameters unnecessary. In response to these issues, Parameter

Efficient Fine-Tuning (PEFT) encompasses a range of techniques to reduce the computational and memory requirements of fine-tuning a pre-trained LLM while still yielding highly effective results. This can be achieved through different techniques, some of which involve reducing the number of parameters to be updated in the pre-trained model. Two categories of methods that use this strategy include adding new task-specific adapter weights or neural network layers to the vanilla model to be fine-tuned (i.e., addition-based methods); or freezing most model parameters and only updating a smaller subset of existing parameters relevant to the task at hand (i.e., specification-based methods). In addition, reparameterization-based methods change the way model parameters are represented or computed to make it more efficient or better suited for a specific task or domain (Zhang et al., 2023).

Relevant to the present study is Low-Rank Adaptation (LoRA; Hu et al., 2021), a reparameterization-based method that also leverages the strengths of specification-based methods. All weights in the original pre-trained feature matrix are frozen, thus preserving the knowledge of the foundational model; the elements to be updated are restricted to a low-rank feature matrix $\Delta W$, which is decomposed into the product of two small feature matrices A and B. Therefore, memory and storage demands are greatly reduced: original parameters are conserved and what is stored are the differences between the original parameters and the fine-tuned weights. This approach is highly efficient, achieving high model quality without introducing latency during inference or reducing the maximum length of inputs. Fine-tuning GPT-3 with LoRA was associated with a 10000 times reduction in the number of features to be trained and required only one-third of memory usage compared to full fine-tuning (Zhang et al., 2023).

Quantum Low-Rank Approximation (QLoRA; Dettmers et al., 2023) further optimizes the process of fine-tuning LLMs in low feature spaces. Before applying LoRA, 4-bit

NormalFloat (NF4) Quantization is applied to the weight matrix of the pre-trained model. NF4 performs quantization based on the quartiles of a normal distribution, which appears to work optimally with the typically normal distribution of LLM weights. An additional level of quantization quantizes the constants from the first quantization themselves, further reducing memory, storage, and time requirements. The authors found that, across different tasks and datasets, the computational advantages of QLoRA do not come at a significant loss in model performance compared to FFT. This technique also seems to work well when the training dataset is small but of high quality (Jha et al., 2023), even when employing smaller LLMs compared to state-of-the-art models (Dettmers et al., 2023).

**Prompt Engineering**

Changing model weights and representations is not the only way to align LLM outputs with users' needs. Prompt engineering helps generative AI understand the intent of queries by crafting prompts that are highly effective, leading to more relevant and accurate responses. Prompt engineering is both art and science and offers no one-size-fits-all solution; however, the literature agrees that a prompt should be concise, logical, explicit, adaptive, and reflective (Lo, 2023), and context, question, format, and examples are recognized as four elements for effective prompting (Jacobsen & Weber, 2023).

The following are popular strategies used in prompt engineering: (1) zero-shot prompting asks the LLM to perform a downstream task that the model was not specifically trained on, without providing any example of what the desired output should look like. This technique tests the model's ability to generalize to new tasks; (2) one-shot and few-shot prompting, on the other hand, consist in crafting a prompt that offers the model one or more examples to learn from; (3) chain of thought (CoT) prompting asks the model to generate an output that does not simply

solve a task, but that also unveils the rationale underlying the response (Kojima et al., 2022). CoT can be implemented by few-shot prompting the model with sample responses that break down the reasoning process behind them, or simply by adding "let's think step by step" at the end of a prompt. Evidence suggests that CoT prompting not only improves model performance but also enhances the model's own reasoning skills, as constructing a logical line of thought reduces the risk of leaping to a conclusion that is linguistically coherent but, in fact, incorrect (Bergmann, 2024).

**Instruction Tuning**

Recently, instruction tuning (Wei et al., 2021) has emerged as a popular SFT technique to tailor pre-trained LLMs for chatbot usage. The main difference between this method and standard SFT lies in the training data, where input-output pairs are enriched with an instructional prompt explicitly directing the model to perform a specific task. Therefore, the training dataset for instruction fine-tuning consists of three elements: (1) Instruction: a natural language input that declares what is the given task that the model is expected to perform (e.g., to continue the previous example, "translate this poem from Italian to English."); (2) Optional context: any supplementary information needed to complete the task (e.g., the Italian poem to be translated); (3) Output: the desired answer that the model is expected to generate in response to the given prompt (i.e., instruction and context) (e.g., the English translation of the given text) (Zhang et al., 2023).

With the incorporation of directive natural language instructions in the training data, this method combines aspects of both fine-tuning a pre-trained LLM and prompt engineering. Combining these strengths in the fine-tuning process should lead to a final product that requires less prompt engineering and fewer few-shot examples to generate outputs aligned with users'

desires (Bergmann, 2024). In addition, this method seems to increase the model's ability to generalize to unseen tasks (Wei et al., 2021). In the present study, our model was trained on one specific task: generating feedback on responses to situational judgment questions based on given criteria. Therefore, our dataset includes only one type of instruction; however, the context of every instruction varies greatly from one sample to another, making generalizability essential to achieve high performance across all items and responses.

Now that the fundamental concepts of fine-tuning and prompt engineering have been elucidated, the following sections will explain how these techniques were applied in the current study and the materials used in our investigations.

**Data**

In this study, the dataset used to train the model and evaluate its performance was provided by Acuity Insights, a testing company that develops, administers, and scores the Casper test. Casper is a SJT adopted by numerous higher education institutions across the world as an admission requirement for programs typically related to healthcare and education. SJTs are a form of assessment where examinees are presented with a scenario describing a particular situation, typically involving a challenge or dilemma, and they are asked how they would react or behave. Casper is designed to assess test takers' social intelligence and professionalism. The Casper items are designed to test one (or more) of ten core skills: collaboration, communication, empathy, equity, ethics, self-awareness, resilience, professionalism, problem-solving, and motivation. Each item is composed of one scenario and three short open-ended questions. After using 30 seconds to reflect, applicants are given up to five minutes to respond and justify their answers. The three responses are scored holistically on a scale from 1 to 9 by human raters, based on guidelines that define themes and qualities that are expected to emerge in each scenario.

Responses are scored on their content, disregarding other aspects such as grammar and writing style. Casper scenarios can be presented both in a text format and through short videos, and answers can be typed out or video-record; however, the present study only included English text-based items that required applicants to type out their response.

Casper does not require any technical knowledge, as it is designed to purely assess soft skills. Due to the nature of the test, scenarios are very diverse (e.g., being stranded on an island with strangers, navigating issues in the workplace, and reflecting on a challenging experience from one's past), and there are no strictly right or wrong answers. Therefore, responses are very unique both across items and between applicants answering the same item, scoring guidelines are not strict evaluation criteria, and it is not possible to develop a single template representative of an ideal response against which all other responses can be compared to give feedback.

The dataset used in the current study contains 211,058 responses to 103 unique text-based scenarios. It discloses the scenario and related questions, applicants' responses and scores, the soft skills assessed for, and two sets of information to guide scoring. The first of this information is the guiding background: it generally consists of a long paragraph delving into the meaning of the focal skills assessed by the item, how they relate to the scenario, and how they are supposed to emerge in the applicant's response. In addition, guiding questions summarize the key concepts expressed in the guiding background in three or four brief questions in the form of "Did the applicant demonstrate [skill]/reflect on [issue]/take [topic] into consideration?". All these elements were used, in their original or in a revised form, to fine-tune the pre-trained LLMs for AFG through multiple training iterations. Figures 1 and 2 show the distributions of scores and skills assessed in the dataset, respectively.

**Figure 1**

*Distribution of Scores in the Full Casper Dataset (N = 211,058)*

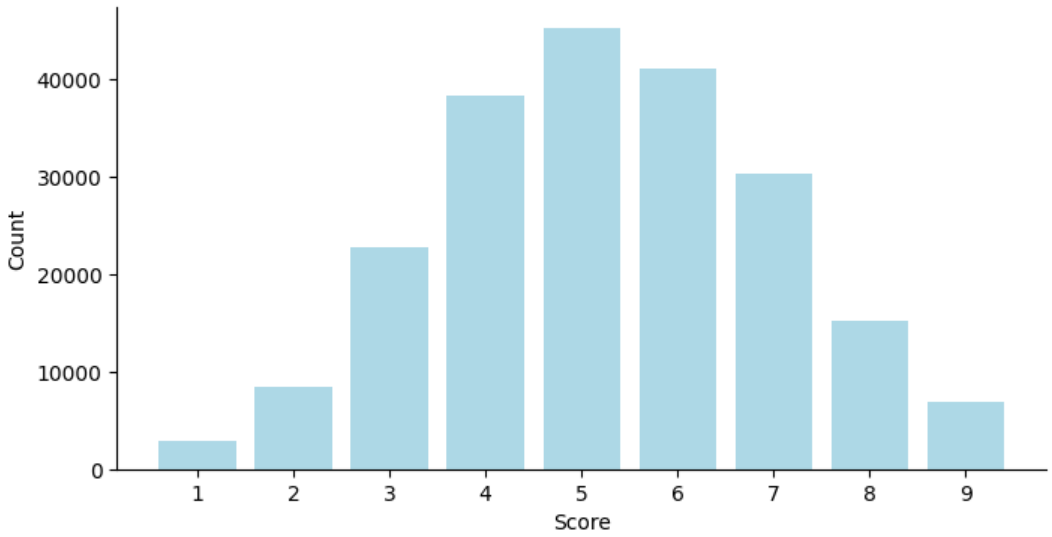
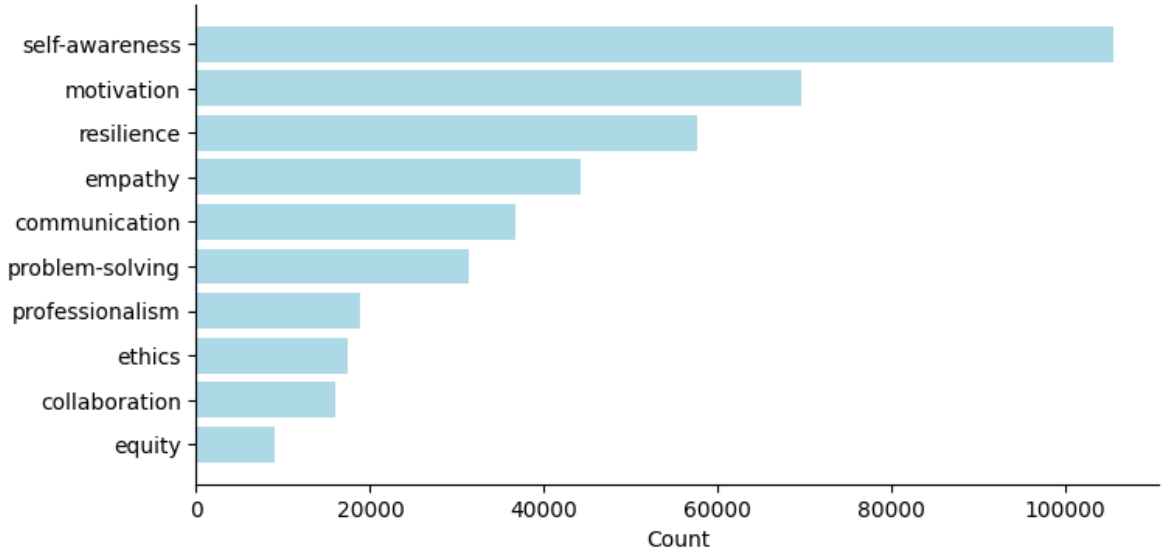



**Figure 2**

*Distribution of Skills Assessed for in the Full Casper Dataset (N = 211,058)*




*Note*. *N* indicates the number of samples in the dataset. For each sample, the item generally measures two essential skills; both are included in the count.

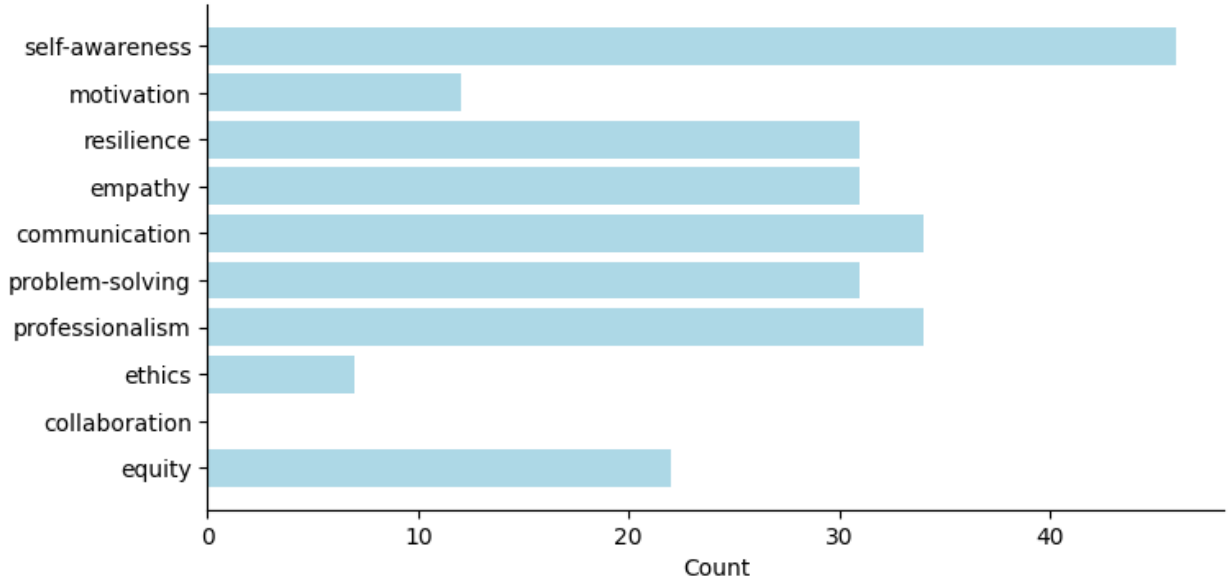
**Training Set Development for Instruction Fine-Tuning**

Supervised fine-tuning requires labeled data to provide the foundational LLM with examples of outputs that best match the user's expectations and needs. For the present study, no example of what “ideal feedback” looks like for Acuity Insight was provided, and thus, the training dataset had to be built from scratch. In general, the larger the training data, the more opportunities for learning are given to the model. However, recent research shows that a smaller high-quality dataset is sufficient to achieve good performance (Jha et al., 2023). While the ideal number of examples varies based on the specific model and use case, OpenAI claims that 50-100 examples are generally sufficient to start seeing a clear improvement in the performance of their GPT models (OpenAI, n.d.). Therefore, we initially developed a set of 100 feedback messages for 100 responses to 12 different scenarios. Additional examples were added to the training data later on, as will be presented below, and the final training set included 124 examples. Casper items were randomly selected to represent a broad range of skills (see Figure 3) and the full range of scores (see Figure 4), in an attempt to help the model learn to give feedback across the full spectrum of competencies and performance levels.

Table 1 outlines in detail how many examples for each score and for each scenario were used for training. Feedback was created in an earnest attempt to incorporate in each message the characteristics of good feedback identified in the literature, with the intention to help applicants improve their performance on the test. Every message was written in the second person (Prins et al., 2006), striving to maintain a supportive tone and balance positive and negative aspects in the response (Hill et al., 2021), with an effort to provide actionable suggestions, to incorporate the evaluation guidelines in the feedback, and to include unique aspects of each response (Hattie and Timperley, 2007).

**Figure 3**

*Distribution of Skills Assessed for in the final Instruction-tuning Dataset (N = 124)*




*Note*. *N* indicates the number of examples in the dataset. For each example, the item generally measures two essential skills; both are included in the count.

**Figure 4**

*Distribution of Scores in the final Instruction-Tuning Dataset (N = 124)*

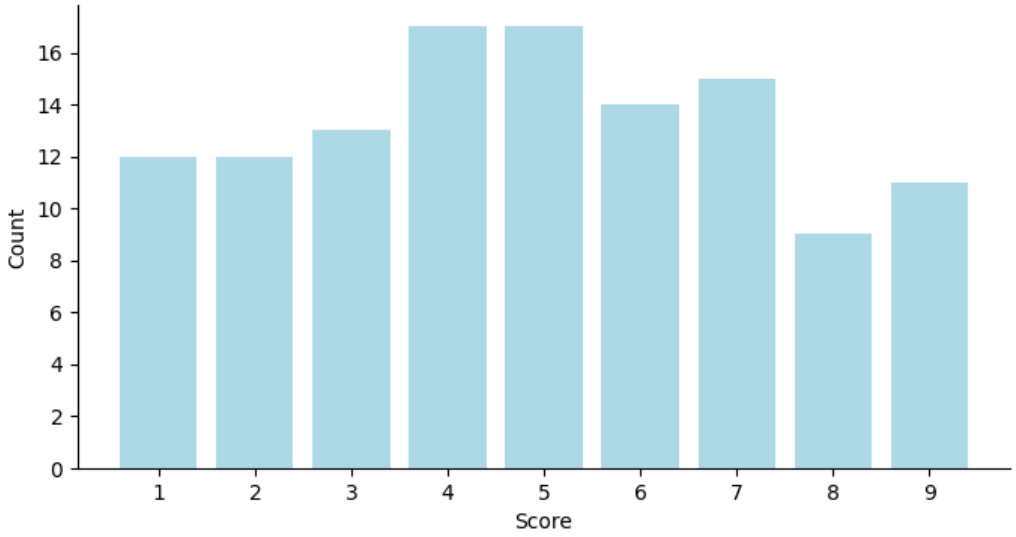


**Table 1**

*Distribution of Examples by Score and Scenario in the Training Dataset*

| | 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 |

**Pre-trained LLMs**

This section introduces the characteristics of the two LLMs that were experimented with in the process of developing the final AFG model. When this project began in September 2023, OpenAI's GPTs maintained their reputation as state-of-the-art models, and Llama-2 (Touvron et al., 2023) gained popularity as a valid open-source competitor. Both models are built on the

decoder-only Transformer architecture and the self-attention mechanisms, boosting their ability to represent contextual relationships, and making these models promising candidates to assist in educational tasks (Kasneci et al., 2023).

***Llama-2-7b***

Llama-2 is a family of pretrained LLMs developed by GenAI and Meta available in three versions, ranging from seven (7B) to 70 billion (70B) parameters. The starting code and weights for the three Llama-2 models were publicly released for research and commercial use in February 2023 (Meta, n.d.). Llama-2 aims to offer capable LLMs without relying on the hundreds of billions of parameters of massive proprietary models, as such dimensionality is associated with computational and infrastructure costs that can be prohibitive for smaller organizations and researchers. Adopting smaller open-source LLMs as the foundation for one's own fine-tuned model offers long-term cost-efficiency compared to massive closed-sourced models; it also reduces privacy risks and concerns, as proprietary data can be used for fine-tuning without being shared with a commercial server and potentially used for future training of commercial models (Bergmann, 2023).

The foundational Llama models were pre-trained on a mix of publicly available online data from a variety of sources, including public sites containing personal information about private individuals. To lower the risk of teaching the model non-factual content, and thus increase the risk of hallucinations, the representation of factual sources was increased during pre-training. Overall, the model was trained on two trillion tokens, and its context length (i.e., the maximum length of the input sequence that the model can handle) was double that of its previous version, reaching 4096 tokens, generally corresponding to 30000 words or six pages of text. Successful examples of fine-tuning the smaller versions of Llama-2 can be found in Alpaca (Taori et al., 2023) and Vicuna (LMSYS Org, 2023): both models achieved high performance

with an investment of only 600 and 300 USD, respectively. However, in these instances, the training data was extremely large, one including 52 million examples, and the other 70 million.

### *GPT-3.5-Turbo-1025*

While the fixed set-up costs for open-source LLMs are convenient in the long term and for high volume traffic, closed-source solutions, with their state-of-the-art capabilities and ease of set-up and maintenance, can still represent a smart financial choice for startups and small-scale projects. Consequently, after the results we achieved with Llama were deemed unsatisfactory and wanting to avoid generating an excessive amount of synthetic data, we turned to GPT-3.5-turbo as a foundational model for fine-tuning.

Since the first release of ChatGPT in November 2022, OpenAI's models have set the standards for state-of-the-art performance. While the details of these models' architectures are not openly divulged, multiple sources suggest that GPT-3 had about 175 billion parameters and GPT-4 is voiced to contain at least one trillion parameters (Albergotti, 2023). GPT-3.5-turbo is an enhanced version of GPT-3 and -3.5, developed to balance performance and efficiency. It is estimated to count around 20 billion parameters (Wodecki, 2023), making it a very affordable solution, especially when compared with the current pricing for the much larger GPT-4 and -4-turbo. GPT-3.5-turbo, version 1025, which was used in the present study, was updated to achieve higher accuracy at responding in a requested format, and, like Llama, the maximum length of its outputs reaches 4096 tokens. The foundational model is available for fine-tuning through the OpenAI API.

## Instruction Fine-Tuning Iterations

These foundational LLMs were fine-tuned for AFG using the instruction-tuning approach. Once personalized feedback messages were hand-crafted for 100 applicant responses, as outlined above, these examples were used as outputs to build the instruction tuning dataset. In

this approach, the instruction itself is a core element of the training data, as it overtly declares the task to complete, and it provides the model with direction on how to complete it. Therefore, in an attempt to achieve high performance, the formulation of the instruction was tweaked in the different iterations of our SFT. The following sections present the specific methods and steps taken to build the final version of our fine-tuned model. All the coding to prepare and launch fine-tuning jobs was performed in Python.

***Fine-Tuning Llama-2-7B***

Given the advantages of open-source LLMs, we first attempted to generate high-quality automatic feedback using Llama-2-7B. To reduce computational demands, we leveraged the PEFT technique of QLoRA, using tools from the HuggingFace libraries PEFT, bitsandbytes, and trl. After tokenization and applying EOS tokens for padding, the base model was quantized using NF4 quantization. Later, LoRA added low-rank adapter weights to the model. All hyperparameters were left to the default settings. The SFT Trainer was created by feeding it the model, the instruction dataset in a .json format, the LoRA configuration, the tokenizer, and the training parameters. Lastly, LoRA weights were merged with the foundational model and the fine-tuned model thus obtained was stored. Thanks to the PEFT approach and due to the small training data available, the training process was very fast (< 30 minutes). The fine-tuned model was then used to generate inferences on applicants' responses unseen during training randomly selected from the Casper dataset.

This process was repeated for three separate training attempts, the only difference being in the instruction of the training data. Table 2 summarizes the structure of the training data for each fine-tuning iteration. An unsystematic qualitative evaluation was carried out after each training attempt to gauge the quality of model inference, compare the outputs obtained from one iteration to the next, and determine how to move forward.

**Table 2**

*Structure of the Instruction Dataset (N = 100) for each Training Iteration Run on Llama-2-7B.*

*Differences Between the Instructions are Highlighted in Bold*

| Fine-Tuned Model | Structure of the Training Data |
|---|---|
| Llama-AFG1 | Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.<br><br>Instruction: Generate feedback for the answer to the following questions: [questions] + Base your feedback on the following criteria: this scenario is meant to assess for [primary skill] + [secondary skill] + [**guiding background**] + Disregard spelling, grammar and style.<br><br>Context: [response] + This answer got a score of [score] out of 9.<br><br>Output: [feedback] |
| Llama-AFG2 | Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.<br><br>Instruction: Generate feedback for the answer to the following questions: [questions] + Base your feedback on the following criteria: this scenario is meant to assess for [primary skill] + [secondary skill] + [**guiding questions**] + Disregard spelling, grammar and style.<br><br>Context: [response] + This answer got a score of [score] out of 9.<br><br>Output: [feedback] |
| Llama-AFG3 | Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.<br><br>Instruction: Generate feedback for the answer to the following questions: [questions] + Base your feedback on the following criteria: this scenario is meant to assess for [primary skill] + [secondary skill] + [**guiding summary**] + Disregard spelling, grammar and style.<br><br>Context: [response] + This answer got a score of [score] out of 9.<br><br>Output: [feedback] |

### *Fine-Tuning GPT-3.5-Turbo-1025*

Since the GPT models are proprietary models owned by OpenAI, no specific information about the techniques used for fine-tuning is available; however, based on statements from Azure CTO Mark Russinovich, the fine-tuning of GPT models also seems to leverage PEFT LoRA to reduce memory requirements (Microsoft Mechanics, 2023). Before requesting a fine-tuning job, the training data had to be adapted to match the format required from the OpenAI API. Thus, each example in the dataset was transformed into a chat-style conversation between the user and the system, where the instruction and context get combined in the user prompt. Table 3 summarizes the structure and differences in the training data between the four fine-tuning interactions that lead to the final model.

## Evaluation

LLMs outputs can be evaluated in three macro areas: (1) linguistic quality, (2) information accuracy, and (3) utility (Celikyilmaz et al., 2020). Aspects of linguistic quality, such as grammatical correctness, fluency, and vocabulary, are often evaluated using automatic techniques, employing metrics such as readability score and lexical diversity. However, most of these metrics cannot be computed without a reference text or they are not suitable to evaluate a text when tasks permit significant diversity (Celikyilmaz et al., 2020). Both these issues apply to the present context, where no one right answer exists, responses are extremely diverse, and there is not necessarily an ultimately correct feedback to any response. Therefore, human evaluation remains the gold standard to ascertain whether the outputs generated by LLMs meet the desired qualities (Van Der Lee et al., 2019); in fact, most studies that have used LLMs for AFG have resorted to rubrics as their main tool for model evaluation.

**Table 3**

*Structure and Size of the Instruction Dataset for each Training Iteration Run on GPT-3.5-Turbo-1025. Differences Between the Instructions are Highlighted in Bold*

| Fine-Tuned Model<br><br>(size of the training data) | Structure of the Training Data |
|---|---|
| GPT-AFG1<br><br>(*N* = 100)<br><br>_____<br><br>GPT-AFG2<br><br>(*N* = 106, **more low scores examples**) | {"messages": [<br><br>{"role": "system", "content": "You are a tutor that generates feedback on responses to situational judgment items based on provided criteria."},<br><br>{"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."},<br><br>{"role": "assistant", "content": "[feedback]"}<br><br>]} |
| GPT-AFG3<br><br>(*N* = 106)<br><br>_____<br><br>GPT-AGF4<br><br>(*N* = 124, **additional examples breaking down the reasoning process. Augmented with the support of GPT-AFG3**) | {"messages": [<br><br>{"role": "system", "content": "You are a tutor that generates feedback on responses to situational judgment items based on provided criteria."},<br><br>{"role": "user", "content": " 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."},<br><br>{"role": "assistant", "content": "[feedback]"}<br><br>]} |

*Note*. *N* indicates the size of the instruction-tuning dataset

To assess how well our GPT-AFG4 model can generate outputs that meet the qualities of effective feedback identified in Chapters 1 and 2, we developed a scoring rubric, and we involved two independent judges as well as Casper experts to evaluate generations giving feedback on Casper responses unseen during training. Moreover, to build a more comprehensive picture of model performance, we ran a small-scale study that allowed participants to interact with the model as test takes, get immediate feedback on their responses, and express their satisfaction with the model output through a survey.

### *Rubric Evaluation*

Our rubric evaluated the generated feedback on the eight criteria presented in Table 4. The rubric was used to evaluate 59 feedback messages generated by GPT-AFG4 on unseen Casper responses. These were randomly selected from the dataset provided by Acuity Insights to obtain a good representation of all scores: six and five samples, respectively, were randomly selected for each score between 1 and 5, and 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 model seemed to have more difficulties in these instances, which are also the ones that might be most in need of feedback. About one-quarter (14) of the selected samples were responses to scenarios used for training.

**Table 4**

*Scoring Rubric for the Internal Evaluation of Model Performance*

| | Criteria | Scoring Categories | | | |
|---|---|---|---|---|---|
| Structure-related criteria | Linguistic Quality | Correct | Spelling error | Syntax error | Semantic error |
| | Factuality | Pertinent | Misinterprets scenario | Misinterprets response | Misinterprets both |
| | Personalization | Personalized | Generic | | |
| | Actionability | Actionable, offers one or more suggestions | Not actionable, no suggestion given | | |
| | Affective Tone | Completely positive | Balanced | Completely negative | |
| | Second Person | Talks to the applicant directly | Does not talk to the applicant directly | | |
| Content-relate criteria | Criteria-based | Not aligned with evaluation guidelines | Somewhat aligned with evaluation guidelines | Aligned | Highly aligned with evaluation guidelines |
| | Focus/Content coverage | Does not raise any valuable points | Incomplete (raises one or more valuable points, but leaves out other important ones) | Complete but unfocused (raises most relevant points, but should have stressed a different aspect of the response to improve) | Complete and focused |

*Note*. Structure-related criteria were evaluated by two independent judges; content-related criteria were evaluated by two Casper experts.

The six structure-related criteria considered in the rubric are quite objective (e.g., whether the feedback contains any grammatical error or provides any suggestions). Therefore, the author of this thesis and a volunteer undergraduate student who received training on the task acted as two independent judges in the evaluation of these aspects. After 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. Due to the limited sample size and high consistency between judges, inter-rater agreement is reported as the percentage of agreement in Table 5. Overall, agreement ranged between 82.6% and 100%, indicating near-perfect agreement between the two raters. This is not surprising, given that, as anticipated, the categories evaluated by the internal raters consist of rather objective qualities.

**Table 5**

*Inter-rater Agreement Between Two Judges Evaluating Samples on Structure-related Criteria*

| Criterion | Inter-rater Agreement |
|---|---|
| Linguistic quality | 82.60% |
| Factuality | 86.95% |
| Personalization | 95.65% |
| Actionability | 100% |
| Valence | 95.65% |
| Second person | 100% |

*Note*. Agreement measured on 23 independently evaluated samples and reported as a percentage of agreement.

On the other hand, determining whether the feedback is giving suggestions that are likely to improve examinees' performance requires a deeper understanding of the test; for this purpose,

two expert Casper judges were recruited to evaluate the same 59 generations in regard to the following content-related criteria: (1) alignment with the evaluation criteria, and (2) completeness and focus of the feedback. Agreement between the two raters was poor on both criteria ($k = 0.09$ and $k = 0.05$, respectively; McHugh, 2012). This lack of alignment between expert judges demonstrates the difficulty to evaluate these characteristics; however, most times, when judges show disagreement, they selected adjacent categories (e.g., judge 1 evaluating an output as aligned with the evaluation guidelines, and judge 2 evaluating the same output as only somewhat aligned, rather than rating it as not aligned with the evaluation guidelines at all).

### *User Evaluation Survey*

Between May 24 and July 18, 2024, a cross-sectional survey study asked participants to assume the role of Casper test-takers and evaluate GPT-AFG4's performance in offering them feedback. Given the low stake of the exercise for participants and the fact that responses were not scored, there was no follow-up investigation to establish whether the feedback would have led to improvement in their performance. The study used a convenience sampling approach, as no sampling frame was available. Participants were recruited primarily among undergraduate and graduate students, as this is likely the most represented demographic in the population of Casper applicants. However, anyone above the age of 18 was allowed to participate in the study; in fact, it is reasonable to believe that there exists a broad variability in the population, for example, test takers could be students from diverse backgrounds looking to transition into healthcare or teaching, or older applicants wishing to further their education. To reach a larger sample size, during the last two weeks of activity, the study was hosted on Amazon Mechanical Turk (MTurk) and 76 additional responses were collected.

In the study, participants were invited to answer one item randomly selected from five retired Casper scenarios, assessing, in total, seven out of the ten social skills. Upon submitting

their response, they immediately received feedback from our final fine-tuned AFG model, and they were asked to evaluate this output by answering a 12-question survey delivered through the online platform Qualtrics. Before deployment, the survey was piloted with three volunteers using a think-aloud protocol. Table 6 offers an overview of the scales included in the survey (the full survey is reported in Appendix A). Similarly to the rubric criteria, the scales were designed to measure users' perception of linguistic quality, factuality, details, personalization, actionability, and affective tone (i.e., balance of strengths and flaws). Based on suggestions from Van Der Lee et al. (2019), all items used a 6-point Likert-type scale ranging from *"1 = completely disagree"* to *"6 = completely agree"*. Originally, three scales included reversed items, where responses declaring higher agreement indicated lower levels of the latent trait; these were re-coded before data analysis to match the direction of the other items in the scale. However, after merging MTurk responses with altruistic ones, these reverse-coded items showed poor item-rest correlations, likely due to careless responding. After removing these problematic items, the affective scale reached a reliability of $\alpha = .69$, approximating the critical threshold of $\alpha = 0.70$ for acceptable reliability for exploratory purposes (DeVellis & Thorpe, 2021); all other scales, achieved good reliability, with coefficient alpha ranging between 0.80 - 0.87 (Table 6).

Later, the polytomous Partial Credit Model (PCM, Masters, 2016) was used to obtain continuous scores for each of the feedback qualities measured in the survey. The PCM is a polytomous IRT model, where the Rasch model is applied to each pair of adjacent response categories to a group of items measuring the latent trait on an ordinal scale. When interpreting the scale scores so obtained, higher values indicate that the evaluated generation was found to demonstrate higher levels of the latent trait. At the end of the survey, one yes-no question recorded overall satisfaction with feedback, and one open-ended question allowed participants to

leave additional comments. Lastly, two demographic questions collected information about student status and whether English was the participant's first language. This information was collected to test for possible differences in the perception of feedback quality between groups.

**Table 6**

*Overview of Scales included in the Survey for User Evaluation of Automatically Generated Feedback*

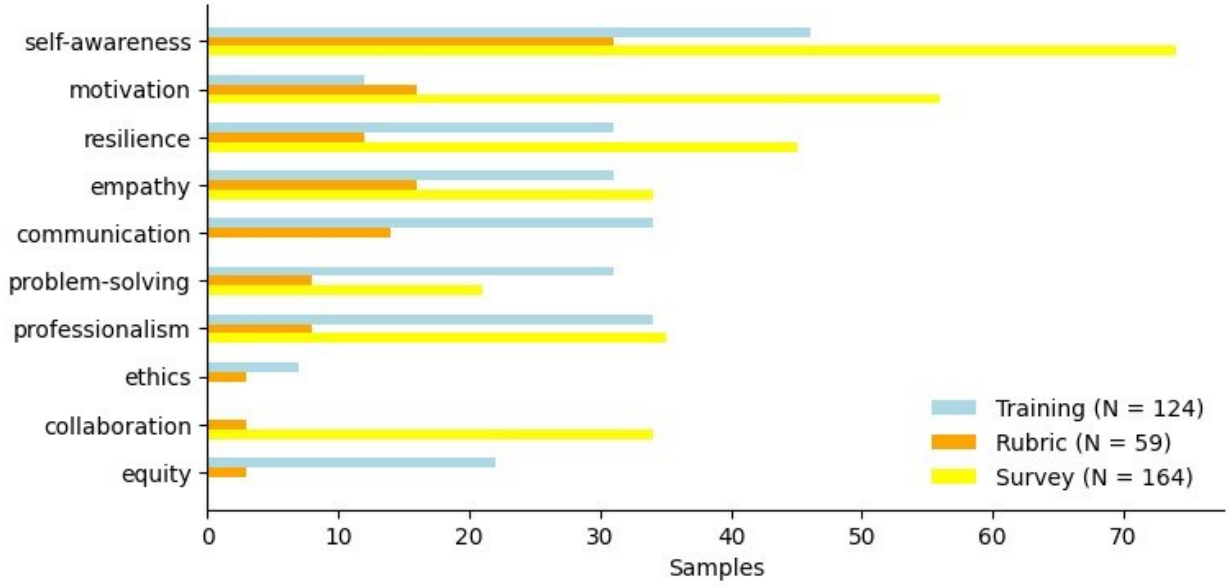
| Construct | Number of Items | Number of Reverse-coded items | Sample Item: The feedback… | Reliability (Coefficient α) |
|---|---|---|---|---|
| Linguistic quality | 6 → 4 | 2 (removed) | is clear | 0.80 |
| Factuality | 4 | 0 | is relevant to my response | 0.80 |
| Details | 4 → 3 | 1 (removed) | is comprehensive | 0.82 |
| Personalization | 4 | 0 | was written specifically for my response | 0.85 |
| Actionability | 5 | 0 | helps me identify how to improve my response | 0.87 |
| Affective tone | 4 → 3 | 1 (removed) | tries to balance positive and negative aspects in my response | 0.69 |

*Note.* All items use a 6-point Likert response scale: 1 = *completely disagree; 2 = disagree; 3 = slightly disagree; 4 = slightly agree; 5 = agree;* 6 = *completely agree*.

After eliminating cases who completed less than 40% of the survey, missingness within items was very low, ranging between 0% and 6.7%. Therefore, the analyses were conducted without performing any imputation. Survey data was preprocessed and analyzed in R (R Core Team, 2023). Overall, the samples used for evaluation through the rubric and the survey study covered the full range of Casper skills (Figure 5).

**Figure 5**

*Distribution of Skills Assessed for in the Samples Used for Fine-Tuning and for Evaluation of Model Performance*




*Note. N* indicates the size of the training and the evaluation samples. For each case, the item generally measures two essential skills. Both primary and secondary skills are included in the count.

# Chapter 4: Results

## Fine-Tuning Iterations on Llama-2-7B

In the first attempt to fine-tune Llama-2-7B (see Table 1), using 100 examples and the guiding background in the instruction led to very poor performance: the model failed to pick up the desired feedback style, and often misinterpreted scenarios and responses. We identified two possible causes: the limited size of the training data, and the lengthiness of the guiding background, as LLMs might struggle to identify the most relevant information in a very long and unfocused prompt (Lo, 2023). Given the high cost of hand-crafting new examples, we first decided to address the second issue: in the second training attempt (Llama-AFG2), the model was given more concise instruction by using the guiding questions in place of the guiding background. This set of guidelines typically asks three or four direct questions that are very focused and highlight the key aspects to look for in the applicant's response.

While some improvement was observed, Llama-AFG2 still presented significant shortfalls in inference, such as summarizing answers without identifying strengths or weaknesses in the response, not offering suggestions for improvement, or not addressing the applicant in the second person. We suspected that this might be caused by the model not gaining enough understanding of the evaluation criteria, due to the very limited context given by the guiding questions. Therefore, in the following fine-tuning iteration, we aimed to strike a balance between the wordiness of the guiding background and the lack of specificity of the guiding questions by combining the two into guiding summaries. Given the large number of scenarios present in the overall dataset, ChatGPT-3.5-turbo was used to automatically summarize each of the 103 pairs of guiding statements. This new set of guidelines was then used in the fine-tuning instruction to train Llama-AFG3 on the same 100 examples. This variation in the prompt seemed to improve performance: the model generated outputs that reflected the style of the training samples, pointed

out shortcomings in the applicant's response while maintaining a positive tone, and offered more suggestions for improvement. However, these results were not consistent, and most generations still presented visible flaws (e.g., misinterpretations and parts of the instruction repeated in the output).

Some researchers suggest that open-source models are still less capable than the state-of-the-art GPT (Roumeliotis et al., 2024), and successful examples of fine-tuning Llama used an extremely large training dataset (e.g., Alpaca - 52 million examples (Taori et al., 2023); Vicuna - 70 million examples (LMSYS Org, 2023)). Wanting all feedback messages to be manually written or checked, it was unfeasible to achieve a similar training size in the present study. Therefore, finding the results obtained from fine-tuning Llama-2-7B too far from the desired outputs, we turned to GPT-3.5-turbo-1025 as another possible foundational LLM for our AFG model.

**Fine-Tuning Iterations on GPT-3.5-Turbo-1025**

GPT-AFG1 was trained using the same prompts and the same 100 training examples as Llama-AFG3 (see Tables 1 and 2). Comparing each model's feedback generated for the same applicant responses, GPT performance was found to be visibly superior. Upon further inspection, we found that the model largely generated pretty good feedback for responses that received middle and high scores but struggled to do the same for low-score responses. Therefore, for the following fine-tuning job, the training data was augmented to include six more examples of feedback on responses that received the lowest possible score (i.e., 1). GPT-AFG2 was fine-tuned on this dataset, without any changes to the prompt structure. Feedback messages generated by this model (five for each score between one and three) were compared to those generated by GPT-AFG1 on the same 15 responses; it was concluded 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 prompting technique, asking the model to "think step by step" (Table 2). This approach did not necessarily seem to bring too large of an improvement to the model. For the next and final training iteration, 18 more examples across six scenarios were added to the training set, taking care that each new sample output would explain the relation between the feedback, the response, and the evaluation criteria. To aid in the creation of so many samples, GPT-AFG3 with zero-shot CoT prompting was leveraged to create a first feedback draft, which was then modified to align the examples more closely with the desired output. Overall, only small changes were needed, mostly to integrate the evaluation criteria more explicitly in the feedback or to shift the focus from one suggestion to another. The obtained GPT-AFG4 cost only 1.97 USD to fine-tune, and it was retained as our final model.

**GPT-AFG4 Evaluation**

The following sections report the result of the systematic evaluation of the outputs generated by this model.

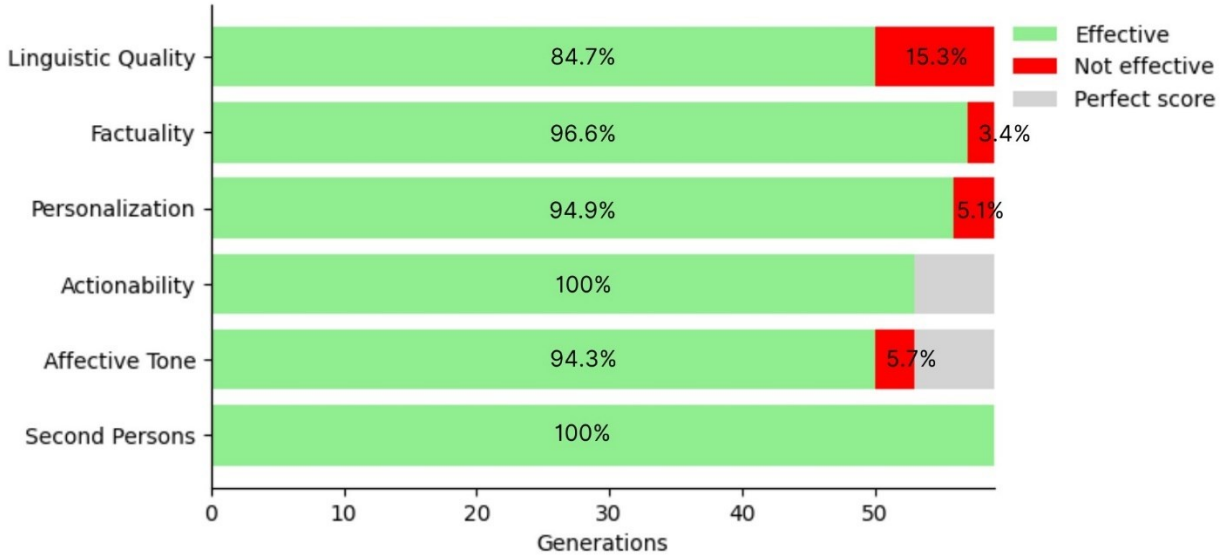
***Rubric Evaluation: Structure-related Criteria***

Evaluation of GPT-AFG4 performance on 59 random generations suggests that during fine-tuning, the model picked up the feedback style and structure observed in the training samples (Figure 6).

*Factuality and linguistic quality.* The model consistently generated feedback pertinent to the context of the scenario and of the response (57/59); however, twice, it misinterpreted the communicative intent of the response. Moreover, in 10 generations, the model committed small linguistic mistakes. In particular, the following errors were detected: six misspelling occurrences, one syntactically incorrect sentence, one sample presenting both spelling and syntactic problems, and one instance of inappropriate semantic use of a word.

**Figure 6**

*The Proportion of Automatically Generated Feedback that Meets the Qualities of Effective Feedback for the Structure-related Criteria Considered in the Rubric (N = 59)*




*Note.* For actionability and affective tone, perfect scores responses are shown separately because in the training examples feedback on answers that achieved the highest score was always completely positive and no suggestions for improvement were offered.

*Personalization and actionability*. GPT-AFG4 generated personalized feedback messages in 56 out of the 59 evaluated samples. The remaining three feedback messages were found to be quite generic, as they did not reference specific elements of the applicant's response and offered advice that would apply to any response. This happened for responses that got both low (2, 3) and high (7) scores. With the exception of feedback messages commenting on responses that achieved the highest score, all generations were evaluated as actionable, as they included suggestions for improvement.

*Affective tone*. All outputs addressed the applicant in the second person, and in almost the totality of instances (50 out of 59), the feedback was found to point out both positive and

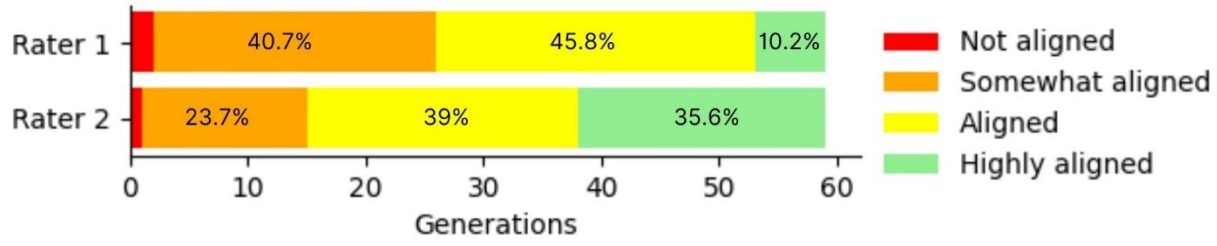
negative aspects of the response. Of the remaining nine cases, six were responses that have been awarded a perfect score (9/9) and were evaluated as completely positive, one was a response that received a score of 8 and its feedback was perceived as completely positive, and two were responses that received scores of 1 and 3 and their feedback was perceived as only pointing out flaws. Overall, 72.9% (43/59) of instances met all structural criteria of effective feedback. Fisher's exact test did not identify a significant difference in the frequency of effective vs flawed feedback between output generated on responses to scenarios seen or unseen during training ($p = 0.48$).

***Rubric Evaluation: Content-related Criteria***

Despite the poor alignment between the two independent evaluations, both judges 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 one or two generations were flagged as not at all aligned with the evaluation criteria (Figure 7). Regarding the completeness and focus of the feedback, both judges found more variability in model performance, each finding only a quarter of generations to be complete and focused (Figure 8).

**Figure 7**

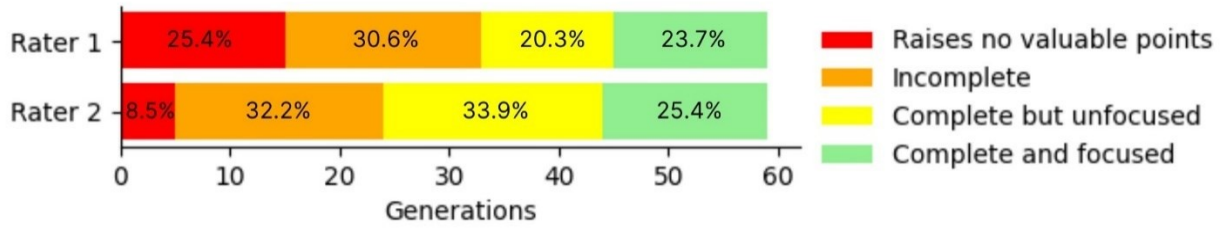
*Casper Experts Evaluation of how Well the Automatically Generated Feedback Aligns with Evaluation Guidelines (N = 59)*




*Note*. Evaluations for both raters are reported due to poor inter-rater agreement. Red: 3.3%, 1.7%, respectively.

**Figure 8**

*Casper Experts Evaluation of Completeness and Focus of the Automatically Generated Feedback (N = 59)*




*Note*. Evaluations for both raters are reported due to poor inter-rater agreement.

***User Evaluation Survey***

After removing cases with excessive missing values, 164 survey responses were retained for analysis. The majority of participants (67.1%) were students or recent graduates, and 17.1% identified as ESL speakers. Each of the five Casper items was answered between 27.4 and 12.8% of the time. The vast majority of respondents (84.8%) expressed satisfaction with the feedback they received. Table 7 summarizes descriptive statistics for each scale score derived from the survey items using PCM. Upon visual inspection, overall scores for each construct present an approximately symmetrical distribution, with scores fairly evenly spread around the mean. However, all scales present at least one outlier in the lower tail of the distribution. Frequency plots for each scale are reported in Appendix B.

*Feedback qualities*. Observing the frequency of responses to individual items, over 70% of respondents agreed or strongly agreed that the feedback was clear, grammatically correct, easy to read, and flowing in a logical order. Similarly, 70% of respondents agreed or strongly agreed that the feedback was pertinent to their response. About 60% of respondents agreed or strongly agreed that the feedback was written specifically for their response. 78.7% of respondents agreed, at least slightly, that the feedback would help them improve their response. However, in

40.9% of generations the model could not satisfy, not even slightly, users' desire for details. Lastly, the model seemed largely able to provide supportive and balanced feedback: 73.2% of participants agreed or strongly agreed that the feedback was supportive, while in fewer than 20% of instances (18.9%) the output was not found, in any measure, to balance positive and negative aspects of feedback. This must include both feedback that was perceived as completely negative and completely positive. Surprisingly, 14 participants (10.4%) claimed that the feedback was not in the second person.

**Table 7**

*Descriptive Statistics for the Scale Scores Obtained from the Items Measuring Characteristics of Effective Feedback Using PCM (N = 164)*

| | Linguistic Quality | Factuality | Details | Personalization | Actionability | Affective Tone |
|---|---|---|---|---|---|---|
| Mean | 0 | 0 | 0 | 0 | 0 | 0 |
| Median | -0.19 | -0.07 | 0.12 | -0.14 | 0.01 | -0.37 |
| SD | 1.66 | 1.65 | 1.83 | 1.85 | 1.55 | 1.52 |
| Minimum | -5.47 | -4.78 | -6.48 | -6.06 | -5.04 | -4.80 |
| Maximum | 3.47 | 3.50 | 3.74 | 3.84 | 3.42 | 3.56 |
| Skewness | 0.22 | 0.40 | -0.22 | 0.02 | 0.18 | 0.36 |
| Kurtosis | 0.30 | 0.15 | 0.14 | 0.48 | 0.61 | 0.59 |

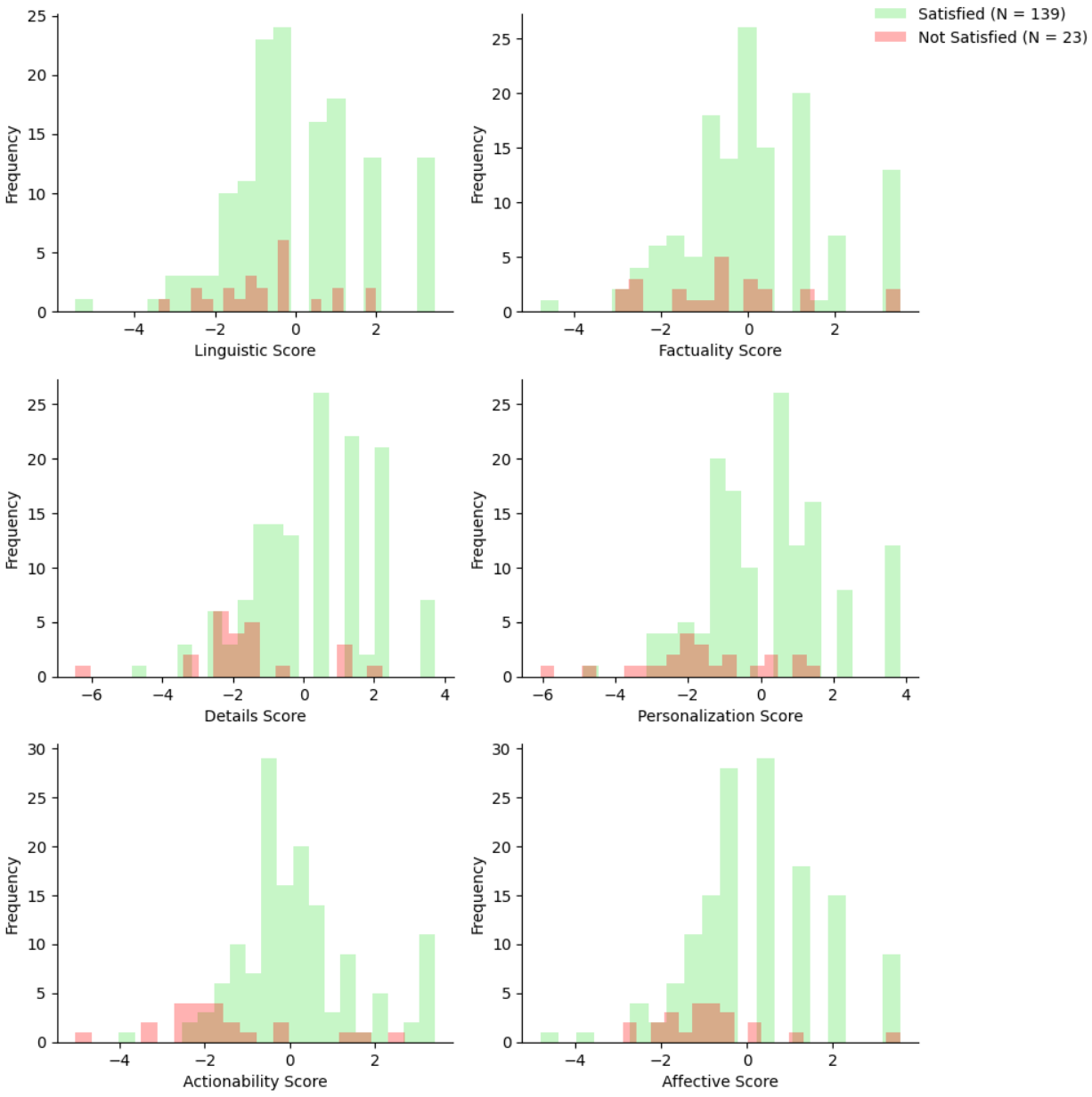
*Note*. SD = Standard deviation

*Group differences*. Fisher's exact tests were performed to investigate differences in satisfaction rates between groups. This method was preferred over a chi-square test of independence due to sparseness in the data, as there were cells in the contingency table showing values smaller than or close to 5 (Renter et al., 2000). The analyses did not find a significant difference in overall satisfaction rates based on ESL status ($p = .55$), student status ($p = .35$), nor scenario answers ($p = .71$).

With only 23 unsatisfied respondents, meaningful analysis to test the association between the scale scores and satisfaction was not feasible: logistic regression typically requires a minimum of 10 to 20 minority cases per predictor to reliably estimate the factors' ability to predict a binary outcome (Harrell, 2015); and conducting multiple analysis or separate t-tests for each factor would increase the risk of type I error (i.e., unduly rejecting the null hypothesis). Therefore, we could not test whether and in what measure the quality of feedback measured in the survey contributed to participants' overall satisfaction with the model's output. However, looking at the distribution of scores for satisfied and unsatisfied respondents, the most notable differences are apparent regarding the perceived level of details, personalization, and actionability (see Figure 9).

**Figure 9**

*Distribution of Feedback Qualities Scores for Satisfied vs. Unsatisfied Survey Respondents (N = 162)*



## Chapter 5: Discussion

Delivering high-quality feedback at scale remains an open challenge in education. Effective AFG systems would not only provide valuable support to student learning (Hattie & Timperley, 2007), but also reduce teacher workload and the associated risk of burnout (Jomuad et al., 2021). Existing AFG systems rarely integrate instructor's knowledge with data-driven techniques and technologies, failing to deliver feedback that is at the same time highly tailor to a task, to a student, and to teacher's needs (Deeva et al., 2021). AFG systems could largely benefit from the capabilities of LLMs. These models' ability to generate text and understand context makes them a promising foundation for the development of AFG systems highly adaptive to different tasks and individual responses. Moreover, their ability to follow instructions could facilitate the integration of teacher's rules and preferences in the generated feedback, making AI a bridge between students' need for personalization and support, and teacher's involvement in the feedback process. Given the recent development of LLMs, only a few studies have explored their application for AFG. Existing evidence mostly uses out-of-the-box proprietary models and relies solely on prompting strategies. This approach shows promising results, but space for improvement remains. Recently introduced fine-tuning techniques allow to customize a pre-trained model for a specific use-case requiring only limited computational resources and training data; this additional training might help the model adapt more closely to the task and desired output characteristics, and it might reduce reliance on educators' ability to craft high-quality prompts (Jacobsen and Weber, 2023). Moreover, in the rapidly evolving landscape of NLP research, open-source LLMs are emerging as valid competitors to the state-of-the-art GPT, achieving high performance while offering cost-effective solutions and increasing data privacy.

Adding to the growing literature on the use of LLMs for AFG, for the first time, the present study reports observations and results from fine-tuning both proprietary (i.e., ChatGPT-3.5-turbo) and open-source (i.e., Llama-2-7B) LLMs for AFG on short open-ended questions. In particular, we tasked the model to generate feedback on responses to a high-stakes SJT measuring social intelligence skills, asking the LLM to generalize widely across items and responses based on given criteria.

**Fine-tuning Open- and Closed-source LLMs for AFG**

Similar to Roumeliotis et al. (2024), observations from our study suggest that fine-tuning GPT for AFG leads to better results compared to further training Llama-2 using the same prompt and number of examples. The difference in performance between the two models was quite striking; however, this conclusion was based only on a qualitative comparison on a small set of generations and using only the smallest version of Llama-2 (i.e., 7B). Further research should conduct a more systematic comparison between open- and closed-source models, including the larger version of the open-source LLM as a foundational model for fine-tuning, as more parameters are generally associated with better performance. However, it is possible that the poor results obtained with Llama-2-7B were due not to the small number of parameters, but to the limited size of the fine-tuning dataset. In fact, successful examples of fine-tuning the smaller versions of Llama do exist, but they used an extremely large training dataset (e.g., Alpaca: Llama-2-7B fine-tuned on 52 million examples (Taori et al., 2023); Vicuna: Llama-2-13B fine-tuned on 70 million examples (LMSYS Org, 2023)). Moreover, a similar conclusion was reached by Roumeliotis et al. (2024), who fine-tuned GPT and the largest available version of Llama-2 on a training set over 28 times larger than ours to predict star ratings from e-commerce reviews. Evaluation of model performance indicated that the fine-tuned GPT outperformed the fine-tuned

Llama-2-70B, as well as both base models, on all classification metrics. The higher success of GPT is attributed to its larger and more diverse training corpus and its more sophisticated architecture compared to Llama-2, as these factors grant GPT broader and more nuanced language understanding capabilities. Nonetheless, given the benefits to costs and data privacy associated with smaller non-commercial models, leveraging open-source LLMs for AFG still warrants further investigation, especially as larger open-source models will likely soon become available (e.g., LLAMA-3.1-405B, Simplifyai, 2024). Therefore, we encourage future research to fine-tune closed-sourced LLMs using more parameters, bigger datasets, and other promising foundational models, such as Mistral (Jiang et al., 2023). These models could be particularly desirable when items or responses could contain highly sensitive information, such as clinical cases in medical education.

This study is also a further testament to the importance of crafting concise, directive, and reflective prompts (Jacobsen & Weber, 2023). Both with Llama-2 and GPT, changing the instructions given to the model from one fine-tuning iteration to the next led to improvements in model performance. During our earliest training iterations, we noticed that the model struggled to provide feedback to responses that were awarded low scores. These answers might be more challenging to provide feedback on because they are sometimes very short and poorly articulated, or because there might be too many aspects that could be touched upon. A similar observation occurred in Steiss et al. (2023) where ChatGPT-3.5 often failed to maintain a positive tone and to offer clear and focused directions for improvement for low-score responses. However, in our study, after providing further examples of feedback on low-score responses and improving the prompt, the issue was no longer evident. This suggests that fine-tuning might be

more effective than prompting in helping the model maintain a balanced tone even when test-takers did not perform very well.

### GPT-AFG4: Strengths, Shortcomings, and Directions for Research

Our final model, GPT-AFG4, was rated quite positively both by judges, test 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 expert educators.

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 got a score of 9, it did not provide any suggestions for improvement; this was also the case in the training data, demonstrating that the model learnt that feedback messages on perfect responses were only to point out the ways in which the applicant met the evaluation criteria.

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 with young children who are still developing their reading and writing skills. We hypothesize that these mistakes might be due to the fact that, in the Casper test, given the limited response time and the focus on the content, applicants' answers sometimes contain typos. No preprocessing was applied to correct these before model training; hence, during fine-tuning, they might have introduced some flaws in the model. To avoid this risk,

future research could try to correct grammatical mistakes in the response before feeding it to the model for training, if the system is only to comment on content.

Furthermore, although the model generally provides recommendations to improve a response, these suggestions are often not comprehensive and focused. This was noted by test experts as well as survey respondents. Some participants expressed the wish for more directive suggestions, such as specific examples of what they could have written or examples of high-scoring responses. Although this is comprehensible from a learner's perspective, such detailed suggestions were purposefully not provided in the example feedback messages. Casper is a high-stakes test measuring character traits that emerge from highly personal responses that do not have to meet any standards or templates. If too much direction was to be given to applicants, this might jeopardize the candor of responses and increase the risk of faking. This partial contraposition between learners' desire for detailed, directive feedback and organizations' responsibility to preserve test security should perhaps be kept in mind in the interpretation of survey results, and it is something for testing companies to consider if they decide to move forward in the development of an AFG tool to support applicants in their preparation for the test.

Overall, our findings suggest that the model was able to generalize across different skills and evaluation criteria, and it represents an encouraging result for the scalability of this system: if fine-tuning a pre-trained LLM on a limited number of items representing a broad set of assessment constructs can prudence a model that provides feedback of similar quality on responses to new items, then testing companies and educators would be able to keep using the same tool as their item bank gets updated, as long as the test maintains its core structure and purpose.

Building on these findings, future research should explore what could be achieved when fine-tuning pre-trained LLMs using a larger example dataset developed by expert educators. Involving teachers in this process and integrating their objectives and criteria into training instructions and examples would give the model more and better opportunities to learn and align to instructors' needs, and also increase transparency for educational stakeholders. In addition, the ideal level of creativity allowed to the model might vary based on subject matter and user need. Therefore, similar to Steiss et al. (2023), the future exploration of fine-tuning for AFG could also adjust the temperature hyperparameter to further control randomness in the model's prediction during text generation. Furthermore, for the sake of the privacy of stakeholders' information, further efforts should be directed to the development of effective and efficient AFG systems that leverage open-sourced LLMs. Moreover, in line with the ethical principle of equality, future studies should explore the development of similar AFG models in languages other than English. For example, the Casper test is offered both in English and French, and applicants who opt to take the test in the Francophone language should have the same opportunity as English-speaking test-takers to benefit from feedback.

**Limitations**

To conclude our discussion and provide further context to interpret our results, it is important to acknowledge some limitations of the present study. First, our study could not examine whether fine-tuning is significantly better than prompting for AFG. Research suggests that through fine-tuning, the model only picks up superficial style characteristics, rather than improving its logical reasoning (e.g., Kung & Peng, 2023). Since no comparison between the two approaches was performed, we cannot rule this out, and it is possible that prompting

ChatGPT in a specific way might yield feedback of similar quality as that generated by our fine-tuned GPT-AFG4 model.

Second, considerations about the validity and generalizability of our survey results are due. This is not only because the sample was relatively small and not representative of the population of Casper test takers, but also due to the following reasons. Not all participants were familiar with this high-stakes test; thus, despite the brief introductory explanation, they probably did not know well the skills they should have demonstrated, how to approach the item, and why it was important to thoroughly justify their stance. This might have led to overly simple responses and impacted participants' perception of their (likely not very positive) feedback. For example, two participants were satisfied with their feedback, but they expressed the feeling that the "*AI expected a too specific and too long answer full of details*", and one participant who was not satisfied with their feedback commented: "*some comments I found helpful [...] I do not understand why clarifying my statements are linked with showing empathy, or if showcasing empathy is the goal of the assignment or not*".

Third, participants' responses to the Casper item did not receive a score; thus, unlike the instruction used for fine-tuning, the prompt did not provide the model with a score to orient its feedback, which might have been an important element for the model to produce high-quality outputs. This might be part of the reason why some survey respondents were not completely satisfied with the suggestions they received, or lack thereof. For example, one participant reports: "*I feel as though my response was average-above average. [...] but I felt like there was tons of room for improvement and I was not given suggestions for improvement or specific examples for improvement.*" It is possible that the model did not offer suggestions because, without knowing it from the score, it struggled to identify that that response had indeed areas for improvement. In

addition, a model is only as good as its training data; therefore, it is plausible that some outputs did not satisfy survey participants and did not meet some criteria in the rubric because the model learned to generate feedback from examples written by a graduate student rather than an experienced educator or a Casper test expert.

Lastly, generative AI is a hot topic right now, and everyone has an opinion on it. It is possible that some survey respondents approached the study with strong enthusiasm or skepticism, which might have impacted their evaluation. For example, one participant left a long comment opening with this disclaimer: "*The person reading these responses should keep in mind that I am very negatively disposed towards generative AI because I know people in academia who have been absolutely swamped by the number of academic misconduct cased using AI, leading to severe overwork and burnout*". It will be left to the reader to guess whether they were satisfied with their feedback.

**Conclusion**

The present study offers a concise overview of current research on the automatic generation of educational feedback. It provides an accessible introduction to the fundamental concepts of LLMs and fine-tuning, and reviews existing evidence from the application of these models for AFG. Addressing an area yet unexplored in the literature, we apply fine-tuning for AFG, leveraging both open-source and proprietary LLMs.

Our findings indicate that, while PEFT allowed for efficient tuning of Llama-2-7B, instruction-tuning yielded better results when using OpenAI's GPT-3.5-turbo as the foundational model. However, due the data privacy concerns associated with the use of proprietary models, as more and larger open models become available, future research should keep exploring open-source alternatives.

Overall, fine-tuning GPT yielded largely positive results, especially considering the characteristics of the training data and limitations inherent to the survey sample. Even when the model is instruction-tuned for a particular task, crafting effective prompts remains crucial to the generation of desired outputs. Using a small hand-crafted training dataset covering a broad range of assessment constructs and scores, fine-tuning was largely successful in teaching the model to align its outputs with the desired feedback structure, and, despite the high diversity in scenarios and responses, the model seemed to generalize effectively across items and performance levels.

Further efforts are needed to address remaining shortfalls, both in the output structure and content. For example, model hyperparameters could be adjusted to optimize performance for specific contexts, and data preprocessing or further prompting could remove mistakes observed in our model. Future research should work closely with educators and content experts to build a larger training dataset strongly aligned with their needs and preferences and extend stakeholder involvement to the evaluation of model performance in authentic settings. Lastly, the question of whether the efforts and cost of fine-tuning bring a significant improvement in model performance over prompting out-of-the-box models remains open, and ethical considerations of equity should prompt the development of AFG systems for other tasks and languages other than English.

## References


Acuity Insights. (2024). *Casper technical manual*. https://acuityinsights.com/resource/casper-technical-manual/

Albergotti, R. (2023, March 24). The secret history of Elon Musk, Sam Altman, and OpenAI. *Semafor*. https://www.semafor.com/article/03/24/2023/the-secret-history-of-elon-musk-sam-altman-and-openai

Azaiz, I., Kiesler, N., & Strickroth, S. (2024). Feedback-Generation for programming exercises with GPT-4. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2403.04449

Bergmann, D. (2023, December 19). *What is Llama-2?* IBM. https://www.ibm.com/topics/llama-2#:~:text=Unfortunately%2C%20the%20tech%20giant%20has,%E2%80%9D%20%E2%80%93%20it%20is%20not.%E2%80%9D

Bergmann, D. (2024, April 5). *What is instruction tuning?* IBM. https://www.ibm.com/topics/instruction-tuning#:~:text=Instruction%20tuning%20thus%20helps%20to,behavior%20more%20useful%20and%20predictable.

Boud, D., & Dawson, P. (2021). What feedback literate teachers do: an empirically-derived competency framework. *Assessment and Evaluation in Higher Education/Assessment & Evaluation in Higher Education*, *48*(2), 158–171. https://doi.org/10.1080/02602938.2021.1910928

Carless, D. (2016). Feedback as dialogue. In *Springer eBooks* (pp. 1–6). https://doi.org/10.1007/978-981-287-532-7_389-1

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education. Artificial Intelligence*, *2*, 100027. https://doi.org/10.1016/j.caeai.2021.100027

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1–45. https://doi.org/10.1145/3641289

Code-Davinci. (2023). *I am code: An Artificial Intelligence speaks: Poems*. Back Bay Books.

Colbran, S., Gilding, A., & Colbran, S. (2016). Animation and multiple-choice questions as a formative feedback tool for legal education. *Law Teacher*, *51*(3), 249–273. https://doi.org/10.1080/03069400.2016.1162077

D'Antoni, L., Kini, D., Alur, R., Gulwani, S., Viswanathan, M., & Hartmann, B. (2015). How can automatic feedback help students construct automata? *ACM Transactions on Computer-human Interaction*, *22*(2), 1–24. https://doi.org/10.1145/2723163

Dawson, P., Henderson, M., Mahoney, P., Phillips, M., Ryan, T., Boud, D., & Molloy, E. (2018). What makes for effective feedback: staff and student perspectives. *Assessment and Evaluation in Higher Education/Assessment & Evaluation in Higher Education*, *44*(1), 25–36. https://doi.org/10.1080/02602938.2018.1467877

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

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLORA: Efficient Finetuning of Quantized LLMS. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2305.14314

DeVellis, R. F., & Thorpe, C. T. (2021). *Scale Development: Theory and applications*. Sage publications.

Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C., Chen, W., Yi, J., Zhao, W., Wang, X., Liu, Z., Zheng, H., Chen, J., Liu, Y., Tang, J., Li, J., & Sun, M. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, *5*(3), 220–235. https://doi.org/10.1038/s42256-023-00626-4

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

Ferguson, R., Hoel, T., Scheffel, M., & Drachsler, H. (2016). Guest editorial: Ethics and privacy in learning analytics. Journal of learning analytics, 3(1), 5-15.

Harrell, F. E. (2015). Regression modeling strategies : with applications to linear models, logistic and ordinal regression, and survival analysis. In *Springer eBooks*. http://ci.nii.ac.jp/ncid/BB19450014

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

Hill, J., Berlin, K., Choate, J., Cravens-Brown, L., McKendrick-Calder, L., & Smith, S. (2021). Exploring the emotional responses of undergraduate students to assessment feedback: Implications for Instructors. *Teaching & Learning Inquiry*, *9*(1), 294–316. https://doi.org/10.20343/teachlearninqu.9.1.20

Holdsworth, J. (2024, June 6). *What is NLP (Natural Language Processing)?* IBM. https://www.ibm.com/topics/natural-language-processing

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Chen, W. (2021). LORA: Low-Rank adaptation of Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2106.09685

Iraj, H., Fudge, A., Khan, H., Faulkner, M., Pardo, A., & Kovanović, V. (2021). Narrowing the Feedback Gap: Examining Student Engagement with Personalized and Actionable Feedback Messages. *Journal of Learning Analytics*, *8*(3), 101–116. https://doi.org/10.18608/jla.2021.7184

Jacobsen, L. J., & Weber, K. E. (2023). 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. *Preprint*. https://doi.org/10.31219/osf.io/cr257

Jha, A., Havens, S., Dohmann, J., Trott, A., & Portes, J. (2023). LIMIT: Less is more for instruction tuning across evaluation paradigms. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2311.13133

Jia, Q., Young, M., Xiao, Y., Cui, J., Liu, C., Rashid, P., & Gehringer, E. (2022). Insta-Reviewer: A Data-Driven approach for generating instant feedback on students' project reports. *Zenodo (CERN European Organization for Nuclear Research)*. https://doi.org/10.5281/zenodo.6853099

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., De Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). Mistral 7B. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.06825

Jomuad, P., Leah, M., Cericos, E., Bacus, J., Vallejo, J., Dionio, B., Bazar, J., Cocolan, J., & Clarin, A. (2021). Teachers' workload in relation to burnout and work performance. *International Journal of Educational Policy Research and Review*, *8*(2). https://doi.org/10.15739/ijeprr.21.007

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., . . . Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Keuning, H., Jeuring, J., & Heeren, B. (2018). A Systematic literature review of Automated feedback generation for programming exercises. *ACM Transactions on Computing Education*, *19*(1), 1–43. https://doi.org/10.1145/3231711

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

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2205.11916

Kung, P., & Peng, N. (2023). Do models really learn to follow instructions? An Empirical study of instruction tuning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2305.11383

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education. Artificial Intelligence*, *6*, 100210. https://doi.org/10.1016/j.caeai.2024.100210

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for natural language generation, Translation, and Comprehension. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1910.13461

LMSYS Org. (2023, March 30). *Vicuna: an Open-Source chatbot impressing GPT-4 with 90%* ChatGPT quality*. https://lmsys.org/blog/2023-03-30-vicuna/

Lo, L. S. (2023). The CLEAR path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship*, *49*(4), 102720. https://doi.org/10.1016/j.acalib.2023.102720

Masters, G. N. (2016). Partial credit model. *Encyclopedia of Social Measurement*, *1*, 109–126. https://doi.org/10.1201/9781315374512-10

Matcha, W., Gašević, D., Uzir, N. A., Jovanović, J., & Pardo, A. (2019). Analytics of learning strategies: associations with academic performance and feedback. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. https://doi.org/10.1145/3303772.3303787

Matelsky, J. K., Parodi, F., Liu, T., Lange, R. D., & Kording, K. P. (2023). A large language model-assisted education tool to provide feedback on open-ended responses. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2308.02439

Mazzullo, E., Bulut, O., Wongvorachan, T., & Tan, B. (2023). Learning analytics in the era of large language models. *Analytics*, *2*(4), 877-898. https://doi.org/10.3390/analytics2040046

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 276–282. https://doi.org/10.11613/bm.2012.031

Meta. (n.d.). *Llama 2: Open source, free for research and commercial use*. Retrieved July 22, 2024, from https://llama.meta.com/llama2/

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). 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. https://doi.org/10.1016/j.caeai.2023.100199

Microsoft Mechanics. (2023, May 23). *What runs ChatGPT? Inside Microsoft's AI supercomputer | Featuring Mark Russinovich* [Video]. YouTube. https://www.youtube.com/watch?v=Rk3nTUfRZmo

OpenAI. (n.d.). *Fine-tuning Guide. Preparing your dataset*. Retrieved July 22, 2024, from https://platform.openai.com/docs/guides/fine-tuning/preparing-your-dataset

Pardo, A., Bartimote, K., Shum, S. B., Dawson, S., Gao, J., Gašević, D., Leichtweis, S., Liu, D., Martínez-Maldonado, R., Mirriahi, N., Moskal, A. C. M., Schulte, J., Siemens, G., & Vigentini, L. (2018). OnTask: Delivering Data-Informed, Personalized Learning support actions. *Journal of Learning Analytics*, *5*(3). https://doi.org/10.18608/jla.2018.53.15

Phung, T., Pădurean, V., Singh, A., Brooks, C., Cambronero, J., Gulwani, S., Singla, A., & Soares, G. (2024). Automating Human Tutor-Style Programming Feedback: Leveraging GPT-4 Tutor Model for Hint Generation and GPT-3.5 Student Model for Hint Validation. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. Association for Computing Machinery. https://doi.org/10.1145/3636555.3636846

Prins, F. J., Sluijsmans, D. M. A., & Kirschner, P. A. (2006). Feedback for general practitioners in training: quality, styles, and preferences. *Advances in Health Sciences Education*, *11*(3), 289–303. https://doi.org/10.1007/s10459-005-3250-z

Pu, G., Jain, A., Yin, J., & Kaplan, R. (2023). Empirical analysis of the strengths and weaknesses of PEFT techniques for LLMs. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2304.14999

R Core Team (2023). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ramos-Soto, A., Vazquez-Barreiros, B., Bugarín, A., Gewerc, A., & Barro, S. (2016). Evaluation of a Data-To-Text system for verbalizing a learning Analytics dashboard. *International Journal of Intelligent Systems*, *32*(2), 177–193. https://doi.org/10.1002/int.21835

Renter, D. G., Higgins, J. J., & Sargeant, J. M. (2000). Performance of the Exact and Chi-sqaure test on sparse contingency tables. *Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1253

Renz, A., & Vladova, G. (2021). Reinvigorating the discourse on Human-Centered Artificial intelligence in educational technologies. *Technology Innovation Management Review*, *11*(5), 5–16. https://doi.org/10.22215/timreview/1438

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

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

Scarlatos, A. (2024). Editorial Overview: Special issue on artificial intelligence in education. *Journal of Educational Technology Systems*, *52*(3), 299–300. https://doi.org/10.1177/00472395241236997

Simplifyai. (2024, July 23). The big leak: Meta's LLAMA 3.1 AI model. *simplifyai*. https://simplifyai.in/2024/07/the-big-leak-metas-llama-3-1-ai-model/

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

Storch, N., & Wigglesworth, G. (2010). Learners' processing, uptake, and retention of corrective feedback on writing. *Studies in Second Language Acquisition*, *32*(2), 303–334. https://doi.org/10.1017/s0272263109990532

Süzen, N., Gorban, A. N., Levesley, J., & Mirkes, E. M. (2020). Automatic short answer grading and feedback using text mining methods. *Procedia Computer Science*, *169*, 726–743. https://doi.org/10.1016/j.procs.2020.02.171

Tan, B., Armoush, N., Mazzullo, E., Bulut, O., & Gierl, M. J. (2024). A review of automatic item generation techniques leveraging large language models. *EdArXiv*. https://doi.org/10.35542/osf.io/6d8tj

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023, March 13). *Alpaca: a strong, replicable Instruction-Following model*. Stanford CRFM. https://crfm.stanford.edu/2023/03/13/alpaca.html

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., . . . Scialom, T. (2023). Llama 2: Open

foundation and Fine-Tuned chat models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2307.09288

Van Der Lee, C., Gatt, A., Van Miltenburg, E., Wubben, S., & Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*. https://doi.org/10.18653/v1/w19-8643

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education, 16*(3), 227–265.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1706.03762

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q., V. (2021). Fine-tuned language models are Zero-Shot learners. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2109.01652

Wiggins, G. (2011). Giving students a voice: the power of feedback to improve teaching. *Educational Horizons*, *89*(4), 23–26. https://doi.org/10.1177/0013175x1108900406

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

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

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2023). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, *55*(1), 90–112. https://doi.org/10.1111/bjet.13370

Yenduri, G., Ramalingam, M., Selvi, G. C., Supriya, Y., Srivastava, G., Maddikunta, P. K. R., Raj, G. D., Jhaveri, R. H., Prabadevi, B., Wang, W., Vasilakos, A. V., & Gadekallu, T. R. (2024). GPT (Generative Pre-trained Transformer) – a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE Access*, *12*, 54608–54649. https://doi.org/10.1109/access.2024.3389497

Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., & Wang, G. (2023). Instruction tuning for large language models: A survey. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2308.10792

Zhang, Z., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing*, *36*, 90–102. https://doi.org/10.1016/j.asw.2018.02.004

# Appendices

## Appendix A.

User evaluation survey

**Can AI offer you good feedback?**
**Testing model performance of a GPT fine-tuned for automatic feedback generation**

Welcome to the study on language models for automatic feedback generation. We are investigating how AI can support learners to reach their goals, and we would like to hear your perspective on whether our model can offer you high-quality feedback.

First, you will be asked to answer a short test assessing soft-skills related to social intelligence and professionalism. This should take no longer than 5 minutes, and no technical knowledge is required. After you submit your response, you will immediately receive feedback, and we will ask you a few questions about its quality.

The survey is anonymous and takes about 10-12 minutes to complete. Participation is voluntary and you can leave the test and the survey at any time by closing your browser windows. Responses will be kept confidential.

If you have any concerns or if you are interested in learning more about this project, please contact the Principal Investigator Elisabetta Mazzullo at mazzullo@ualberta.ca or the research supervisor Dr. Okan Bulut at bulut@ualberta.ca.

By participating in the survey, you declare that you have read and agreed to the Participant Consent Form.

Click on the arrow to start the survey.

Follow the link to take a short test. There is no time limit, but you should spend no more than 5 minutes to write your responses. Your response and your feedback will not be stored.

[Link to web app] (single link for all of the five scenarios: each time the webpage is loaded one scenario is randomly selected and displayed)

After you receive feedback, please come back to this page to let us know what you think about it.

Which scenario did you answer to?

- ○ You are stranded on an island with four other people (1)
- ○ You regularly find yourself without any work to do (2)
- ○ You receive a message from a friend you haven't talked to in over a year (3)
- ○ Forgiveness and compassion are always linked (4)
- ○ You are responsible for a team of volunteers (5)

The following section asks you to evaluate the linguistic quality of the feedback you received.

**Indicate your level of agreement with the following statements.**

The feedback is…

| | 1 - Completely disagree (1) | 2 - Disagree (2) | 3 - Slightly disagree (3) | 4 - Slightly agree (4) | 5 - Agree (5) | 6 - Completely agree (6) |
|---|---|---|---|---|---|---|
| clear (1) | ○ | ○ | ○ | ○ | ○ | ○ |
| grammatically correct (2) | ○ | ○ | ○ | ○ | ○ | ○ |
| easy to read (3) | ○ | ○ | ○ | ○ | ○ | ○ |
| repetitive (4) | ○ | ○ | ○ | ○ | ○ | ○ |

| | | | | | | |
|---|---|---|---|---|---|---|
| flowing in a logical order (5) | ○ | ○ | ○ | ○ | ○ | ○ |
| wordy (6) | ○ | ○ | ○ | ○ | ○ | ○ |

The following section asks you to evaluate the relevance and completeness of the feedback you received.

**Indicate your level of agreement with the following statements.**

The feedback is relevant...

| | 1 - Completely disagree (1) | 2 - Disagree (2) | 3 - Slightly disagree (3) | 4 - Slightly agree (4) | 5 - Agree (5) | 6 - Completely agree (6) |
|---|---|---|---|---|---|---|
| to the scenario (1) | ○ | ○ | ○ | ○ | ○ | ○ |
| to the questions (2) | ○ | ○ | ○ | ○ | ○ | ○ |
| to my response (3) | ○ | ○ | ○ | ○ | ○ | ○ |

| | | | | | | |
|---|---|---|---|---|---|---|
| to the scope of assessment (i.e., showcasing relevant soft skills rather than writing style) (4) | ○ | ○ | ○ | ○ | ○ | ○ |

The feedback...

| | 1 - Completely disagree (1) | 2 - Disagree (2) | 3 - Slightly disagree (3) | 4 - Slightly agree (4) | 5 - Agree (5) | 6 - Completely agree (6) |
|---|---|---|---|---|---|---|
| is comprehensive (1) | ○ | ○ | ○ | ○ | ○ | ○ |
| provides an adequate level of details (2) | ○ | ○ | ○ | ○ | ○ | ○ |
| explains why some elements of your response are | ○ | ○ | ○ | ○ | ○ | ○ |

| | | | | | | |
|---|---|---|---|---|---|---|
| adequate or not (3) | | | | | | |
| is too generic (4) | ○ | ○ | ○ | ○ | ○ | ○ |

The following section asks you to evaluate the feedback you received on aspects that can impact its effectiveness.

**Indicate your level of agreement with the following statements.**

The feedback...

| | 1 - Completely disagree (1) | 2 - Disagree (2) | 3 - Slightly disagree (3) | 4 - Slightly agree (4) | 5 - Agree (5) | 6 - Completely agree (6) |
|---|---|---|---|---|---|---|
| is tailored to my response (1) | ○ | ○ | ○ | ○ | ○ | ○ |
| provides suggestions specific to my response (2) | ○ | ○ | ○ | ○ | ○ | ○ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| was written specifically for my response (3) | ○ | ○ | ○ | ○ | ○ | ○ | |
| references elements that are unique to my response (4) | ○ | ○ | ○ | ○ | ○ | ○ | |

The feedback…

| | 1 - Completely disagree (1) | 2 - Disagree (2) | 3 - Slightly disagree (3) | 4 - Slightly agree (4) | 5 - Agree (5) | 6 - Completely agree (6) | Not Applicable (7) |
|---|---|---|---|---|---|---|---|
| helps me identify what I did well in my response (1) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| helps me identify what to improve in my | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| response (2) | | | | | | | |
| helps me identify how to improve my response (3) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| will help me write a better response (4) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| helps me understand what is expected of a good response (5) | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

The following section asks you to evaluate the affective tone of the feedback you received.

**Indicate your level of agreement with the following statements.**

Was the feedback written in the second person? (talking directly to you)

○ Yes (1)

○ No (2)

The feedback...

| | 1 - Completely disagree (1) | 2 - Disagree (2) | 3 - Slightly disagree (3) | 4 - Slightly agree (4) | 5 - Agree (5) | 6 - Completely agree (6) |
|---|---|---|---|---|---|---|
| is supportive (1) | ○ | ○ | ○ | ○ | ○ | ○ |
| makes me feel judged (2) | ○ | ○ | ○ | ○ | ○ | ○ |
| tries to balance positive and negative aspects in my response (3) | ○ | ○ | ○ | ○ | ○ | ○ |
| is encouraging (4) | ○ | ○ | ○ | ○ | ○ | ○ |

Overall, are you satisfied with the feedback you received?

○ Yes (1)

○ No (2)

If the questions in this survey did not allow you to express the reasons for your satisfaction/dissatisfaction, briefly motivate your response in the comment box below (optional):

_____

Before you go, we would like to learn a little bit about you to explore how the feedback experience might differ for different groups.

Are you a student or recent graduate (1 year)?

○ Yes (1)

○ No (2)

Is English your first language?

○ Yes (1)

○ No (2)

This is the end of the study.

All responses are anonymous, so it will not be possible to withdraw your data after submission. Do you wish to submit your responses?

○ Yes, submit my responses (1)

○ No, withdraw my data (2)

Display This Question only if participant select “No, withdraw my data”

Close your browser window to withdraw your data. Your response will not be stored.

We thank you for your time spent taking this survey.

**Appendix B.**

**Figure B1.**

*Distribution of Feedback Qualities Scores in the Sample of Survey Respondents (N = 164)*