ECE 1786 Course Project Final Report: RateMyInterview

WingYan Lee, Melissa Wong, Haobo Xu, Qi Zhang
1988 words, no penalty

I. Introduction

In today's highly competitive job market, standing out during the interview process is more important than ever. To help job seekers improve their performance, this project introduces RateMyInterview - an interview preparation tool designed to improve responses to behavioural interview questions. RateMyInterview presents users with behavioural questions, evaluates their answers, and provides constructive feedback to help refine their responses. Unlike technical questions, behavioural interview questions involve greater variability and require thoughtful, well-structured answers, making them more challenging to prepare for. RateMyInterview empowers job seekers to develop their skills, gain valuable insights, and build the confidence needed to excel in interviews with prospective employers.

II. Illustration / Figure

Figure 1 below shows the overall system architecture of RateMyInterview. The architecture will be described in further detail in the Architecture and Software section of this report.

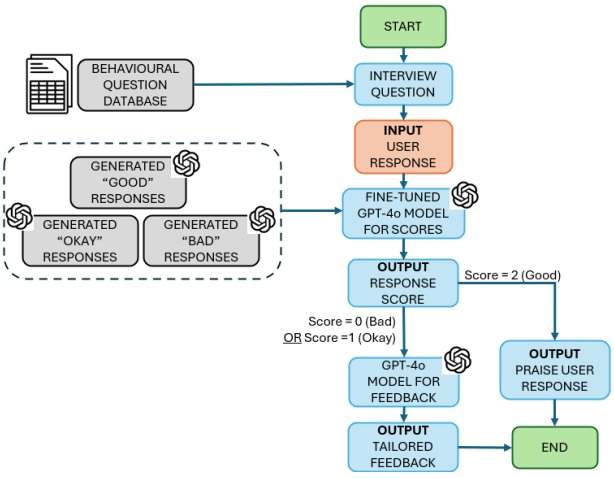

Fig. 1. Overall System Architecture

III. Background & Related Work

In recent years, there have been several studies related to creating interview preparation tools through the use of an LLM. J. M. C J et al. developed the Q&AI Mock Interview Bot which generates interview questions solely based on the user's uploaded resume [1]. The question generation was performed using the BERT / RoBERTa transformer models. To assess response accuracy, the text was summarized using Hugging Face transformers and then passed into a GPT model. In a separate paper, M. Mejias et al. followed a prompt engineering methodology to create a Software Interview Agent called TECH which would simulate an in-person interview using GPT-4o [2]. Through the developed prompt, the application would ask the user a technical coding question, answer questions the user may have, and provide final feedback to the user's answer.

IV. Data

Our team created a semi-synthetic dataset. The dataset consists of 3 classes: good (2), okay (1), and bad (0). They represent the type of interviewee and the quality of answers. 20 distinct behavioral interview questions were selected from various online resources such as Indeed and Novoresume. For each question and for each class, 10 answers were generated by the GPT-4o model with tailored prompts. For each question, we also collected 1 real answer and manually labelled it. After labelling the real data, we found out that the real data only has data from the "good" and "okay" classes. So, we randomly selected 4 questions and manually wrote a "bad" answer for them to create 4 entries of "bad" data in the real dataset. Therefore, the final dataset has 600 generated answers and 24 real answers.

Using our real-life interview experiences, we designed a rubric to help us classify real data and check if all generated data belong to the desired class. The rubric has the following six categories:

Level of detail: checks whether the answer includes highly specific details of action. Outcome: checks whether the answer clearly explains the positive outcome of the action, quantifying results if possible and demonstrating a clear contribution to the success of a project.

Relevance: checks if the response directly addresses the question.

Structure: checks if the response is clearly and logically structured, following the STAR method. The STAR method is a structured approach to answering behavioral interview questions by organizing your response into four parts: Situation, Task, Action, and Result. It helps you clearly describe the context of a challenge, your specific role, the steps you took to address it, and the positive outcome or impact of your actions.

Language: ensures the answer uses professional and appropriate language.

Skill demonstration: ensures the response effectively highlights good soft skills or qualities sought by the employer.

A maximum of 2 and minimum of 0 point will be given to each category based on how much the answer satisfies the criteria of that aspect. We calculated the total points and used a decision range (Good: 9-12; Okay: 5-8; Bad:0-4) to classify answers into the 3 classes. Then, we randomly split the generated dataset by 60:20:20 into fine-tuning (360 pairs), validation (120 pairs), and test (120 pairs) set. We included the 24 real answers into the testing set, which makes it 144 pairs.

Below are sample data for 1 question.

Question: Describe a time when you didn't know how to solve a problem. What did you do to resolve the issue?

Bad (0) answer: Hah, let me tell you about the time I was trying to fix this old, janky coffee machine at work.

...

Okay (1) answer: I remember encountering a problem at work where I wasn't sure how to proceed.

...

Good (2) answer: When I worked in customer support, I used to get tasked with solving problems I knew nothing about.

...

(Please refer to Appendix A for complete samples)

V. Architecture of our system

The architecture of our system is shown in Figure 1.

Our first agent is a GPT-4o model fine-tuned on our synthetic dataset. For a given question & answer pair, it will return a score with a range of 0-2, 0 for bad quality, 1 for okay quality, and 2 for good quality.

Our second agent is a GPT-4o model configured using a prompt. For a given question & answer pair, it will return a tailored feedback to help the user improve the quality of their answer. As is told by the prompt of this model, the generated feedback strictly follows the STAR principle, and will also cover additional suggestions if the user's answer is imperfect in structure (e.g. problematic general structure), vocabulary (e.g. contains casual language), and detail (e.g. vague, incomplete or unrealistic details). The prompt of this model has several early versions, and its final version is defined by our discussion. The definition of this prompt comes from the past interview experience of our team.

VI. Baseline Model & Comparison

Our system outputs both a score to classify responses and feedback to improve the input answer. To measure the success of the scoring part, we used the test dataset and a confusion matrix to evaluate the model performance because the real dataset is imbalanced with majority of answers being class "good". Also, we configured a GPT-4o model with the same prompt as our baseline model. This baseline model will not undergo the fine-tuning process, and will be compared with the fine-tuned model in the later sections.

For the feedback model evaluation, we designed a rubric to manually evaluate the quality of each generated feedback. The rubric has the following 6 categories:

Question relevance: checks if the feedback is relevant to the interview question. Interviewee response relevance: checks if the feedback focuses on key aspects of interviewee's responses.

Actionable feedback: ensures the feedback provides actionable suggestions.

Tone: ensures the tone is professional and encouraging.

STAR: ensures feedback explicitly mention about aligning with the STAR structure.

Positive acknowledgement: checks if feedback highlights strengths in the interviewee's response.

We gave a maximum of 2 and a minimum of 0 point to each category based on the quality, calculated the total points, and used a decision range (Good: 9-12; Okay: 5-8; Bad:0-4) to classify.

VII. Quantitative Results

The performance of the model was evaluated using two datasets: a generated test dataset and a real test dataset. We measured performance using the following metrics:

1) Accuracy:
   Our model achieved an accuracy of 1.0 (100%) on the generated test dataset. However, the accuracy dropped to 0.54 (54%) on the real test dataset. The baseline model achieved an accuracy of 0.64 (64%) on the generated test dataset and an accuracy of 0.75 (75%) on the real test dataset.
   The difference in accuracy highlights the model's overfitting to generated data and its struggle to generalize to real-world inputs. Conversely, the baseline model shows better adaptability to real data but struggles with generated data.
2) Normalized Confusion Matrix:

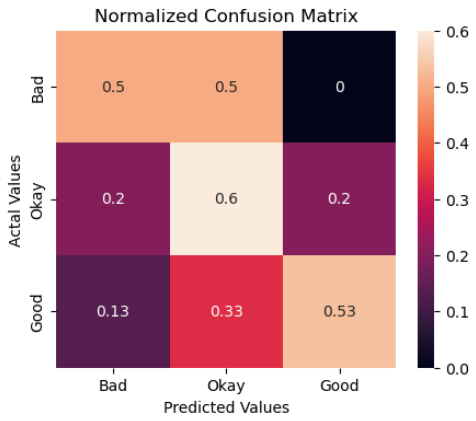

Fig. 2. RateMyInterview Scores on Real Test Dataset

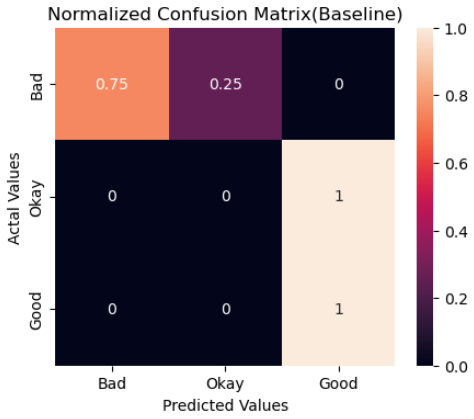

Fig. 3. Baseline Model Scores on Real Test Dataset

For the real dataset, the confusion matrix reveals that our model has difficulty distinguishing between "Bad" and "Okay" responses, as reflected in the off-diagonal values. The baseline model performs better in this regard but struggles with "Okay" classifications.

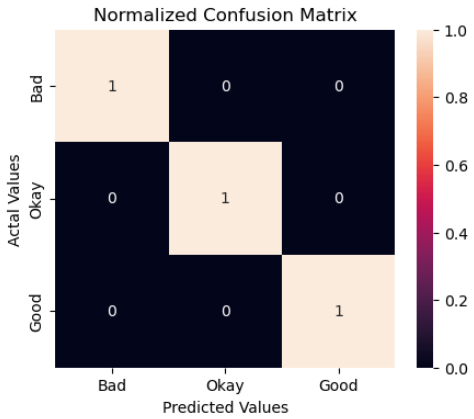

Fig. 4. RateMyInterview Scores on Generated Test Dataset

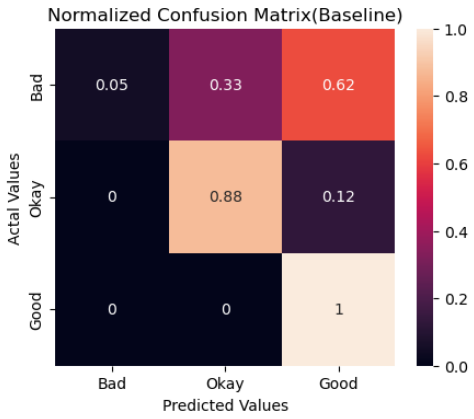

Fig. 5. Baseline Model Scores on Generated Test Dataset

For the generated dataset, the confusion matrix for our model shows perfect alignment between predictions and true labels, with all diagonal values equal to 1. In contrast, the baseline model exhibits classification challenges for generated data.

These effectively highlight both overall performance (accuracy) and specific classification challenges (confusion matrix), providing a comprehensive view of the model's capabilities and limitations.

VIII. Qualitative Results

To evaluate the qualitative performance of our model, we analyzed both its response scores and the quality of the feedback it generated. While the model demonstrated better score prediction accuracy on the generated dataset compared to the real dataset, its feedback quality remained consistently high across all datasets.

As mentioned earlier, we evaluated the feedback using a rubric. The manual evaluation of all feedback in the test set revealed that every instance was rated as good feedback, indicating consistent quality across datasets despite variations in response score accuracy.

Below are sample inputs and outputs.

1) Sample with correct score and good feedback:

Input:

Q: Describe a time when you didn't know how to solve a problem. What did you do to resolve the issue?

A: (Please refer to the Bad (0) example answer in Data section)

Output:

Score: 0,

Feedback: Feedback Based on the STAR Principles - Situation: Your description of the situation is engaging but lacks professionalism and clarity. You might also consider eliminating casual expressions and ensuring that the context of the problem is clear.

Example Improvement: In the office, we relied heavily on a coffee machine due to its constant use providing staff with refreshments. One morning, it suddenly stopped working, affecting everyone's routine.

...

(Please refer to Appendix B for the full feedback)

2) Sample with wrong score (correct score is 2) and good feedback:

Input:

Q: (Same question as above sample)

A: (Please refer to the Good (2) example answer in Data section)

Output:

Score: 1,

Feedback: Feedback based on the STAR principles: - Situation: The situation is well-defined, putting the interviewee in customer support facing an unknown technical issue independently during a shift. However, a bit more context about the stakes or the typical procedures in such scenarios could enhance clarity.

Example Improvement: When I worked in customer support, I was frequently tasked with resolving unfamiliar problems. One evening, as the most senior shift member during a peak customer service time, a customer called in with a technical issue I had never encountered. Due to the high-pressure environment, resolving such issues quickly was crucial to maintaining customer satisfaction and service standards.

...

(Please refer to Appendix C for the full feedback)

IX. Discussion and Learnings

The comparison between the baseline model and the fine-tuned model shows two significant performance gaps. The fine-tuning process drastically increased the model's accuracy on the generated test dataset, but the fine-tuned model has a much worse performance on the test dataset containing real samples.

After further discussion and reflection, we found out several possible causes to this issue. First, our generated samples are too typical, and several types of answers(For example, answers with incomplete detail) are not covered. Thus our fine-tuned model is lacking in knowledge to these answers and will just return a wrong score for them. Secondly, for the fine-tuned model, we stayed with the initial version of the prompt, which is also problematic. If we could develop a better prompt, the model will probably have a better performance on the real dataset. For similar future projects, we would generate more diverse synthetic data to simulate real input variability and refine the prompt for better performance.

X. Individual Contributions

1) WingYan:
   a) Generated "bad" dataset and manually evaluated generated response
   b) Collected the 20 behavioral questions
   c) Collected the real dataset and wrote "bad answer" examples
   d) Manually evaluated 1/4 of the test dataset's generated feedback
   e) Helped create interview response and feedback rubrics
2) Melissa:
   a) Generated "okay" dataset and manually evaluated generated response
   b) Created gradio interface for the presentation demo
   c) Conducted background research on other papers
   d) Manually evaluated 1/4 of the test dataset's generated feedback
   e) Helped create interview response and feedback rubrics
3) Haobo:
   a) Fine-tuned the GPT-4o model to evaluate the answers
   b) Implemented a baseline model for comparison
   c) Configured the GPT-4o model for feedback generation
   d) Manually evaluated 1/4 of the test dataset's generated feedback
4) Qi:
   a) Generated "good" dataset and manually evaluated generated response
   b) Manually evaluated 1/4 of the test dataset's generated feedback
   c) Helped create interview response and feedback rubrics

References

[1] J. M. C J, M. Sabi, M. Benson, G. Baburaj, and S. S, "Q&ai: An ai powered mock interview bot for enhancing the performance of aspiring professionals," in *2024 International Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication, and Computational Intelligence (RAEEUCCI)*, 2024, pp. 1–5.

[2] M. Mejias, Z. Vargas, W. T. Edwards, G. Washington, L. Burge, D.-M. Wilson, and L.-M. Kouaho, "Equity in the preparation of students for software engineering coding interviews: Chatgpt as a mock interviewer," in *2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)*, 2023, pp. 1016–1020.

Appendix

A. Full sample of "Bad", "Okay", and "Good" user answer

Bad (0) answer: Hah, let me tell you about the time I was trying to fix this old, janky coffee machine at work. It was one of those mornings, you know? I pressed all the buttons, went for a slap or two, classic reboot tactics, right? Nada. The thing was deader than my houseplants. So, I was just about to chuck it out the window when I figured, 'Okay, chill. Use your resources, mate.' Hit up Dave from maintenance on chat, he's a legend and knows his way around gadgets. Turns out, he was in the building! So, he strolled over, took one look, and pointed out something was unplugged behind the counter. What a facepalm moment! I guess that taught me: sometimes just asking for help does wonders. And it was way better than trying to become a DIY electrician on a lack-of-caffeine brain. Plus, it saved everyone from those store-bought lattes, which, let's be real, taste like cardboard.

Okay (1) answer: I remember encountering a problem at work where I wasn't sure how to proceed. I felt a bit overwhelmed at first, but I decided to take some time to think about it. I considered different approaches and tried to gather some information. I also talked to a few people to get their perspectives, hoping to get some ideas from them. It took a while, but eventually, I found a way to handle the situation. I guess sometimes I tend to overthink things, which can slow me down a bit.

Good (2) answer: When I worked in customer support, I used to get tasked with solving problems I knew nothing about. One evening, I found myself as the most senior person on shift, despite being relatively new myself. Right then, a customer called with a technical issue that I had never encountered before. So, I needed to resolve the customer's issue and make sure they're satisfied with our support, while at the same time, I had to learn how to fix the problem with no one to help me. I was a little tense but kept my composure and assured the customer that I would find a solution and get back to

them shortly. So, after hanging up, I did some research. I started by checking our internal knowledge base, then I started searching tech forums for similar issues. It took a bit of time, but I was able to piece together the information I needed to troubleshoot the problem. In the end, I called the customer back with a step-by-step solution, which resolved their issue. They were grateful for the help and they were pretty satisfied with how I handled the problem, even though it took longer than it probably should have. I learned how to handle that particular problem like a pro, and I felt a lot more confident at work after that. Now, I feel like I can handle just about any unfamiliar problem that comes my way.

B. Full sample with correct score and good feedback

Input: Q: Describe a time when you didn't know how to solve a problem. What did you do to resolve the issue?

A: Hah, let me tell you about the time I was trying to fix this old, janky coffee machine at work. It was one of those mornings, you know? I pressed all the buttons, went for a slap or two, classic reboot tactics, right? Nada. The thing was deader than my houseplants. So, I was just about to chuck it out the window when I figured, 'Okay, chill. Use your resources, mate.' Hit up Dave from maintenance on chat, he's a legend and knows his way around gadgets. Turns out, he was in the building! So, he strolled over, took one look, and pointed out something was unplugged behind the counter. What a facepalm moment! I guess that taught me: sometimes just asking for help does wonders. And it was way better than trying to become a DIY electrician on a lack-of-caffeine brain. Plus, it saved everyone from those store-bought lattes, which, let's be real, taste like cardboard.

Output:

Score: 0,

Feedback: Feedback Based on the STAR Principles - Situation: Your description of the situation is engaging but lacks professionalism and clarity. You might also consider eliminating casual expressions and ensuring that the context of the problem is clear.

Example Improvement: In the office, we relied heavily on a coffee machine due to its constant use providing staff with refreshments. One morning, it suddenly stopped working, affecting everyone's routine.

- Task: This part wasn't explicitly defined. You should clearly articulate your responsibility or role in this situation.

Example Improvement: As a member of the office support team, I felt it was my responsibility to address the issue swiftly to minimize disruption.

- Action: While you described your actions with humor, it's crucial to maintain a professional tone and clearly outline the steps you took to resolve the issue, including how you leveraged available resources.

Example Improvement: Initially, I attempted basic troubleshooting techniques such as examining the machine's settings. When these efforts were unsuccessful, I decided to consult with Dave from the maintenance team, known for his expertise with office equipment.

- Result: You communicated the result effectively, but aim for a more concise and professional tone. Additionally, highlight what you learned explicitly.

Example Improvement: Dave's insight led to the discovery of an unplugged power cable, swiftly resolving the issue. This experience reinforced the value of collaboration and seeking assistance when needed, preventing unnecessary downtime and maintaining office morale.

By organizing your answer according to the STAR method and using professional language, you can present the experience as both humorous and insightful while demonstrating problem-solving skills effectively."

C. Full sample with wrong score (correct score is 2) and good feedback

Input:

Q: Describe a time when you didn't know how to solve a problem. What did you do to resolve the issue?

A: "When I worked in customer support, I used to get tasked with solving problems I knew nothing about. One evening, I found myself as the most senior person on shift, despite being relatively new myself. Right then, a customer called with a technical issue that I had never encountered before. So, I needed to resolve the customer's issue and make sure they're satisfied with our support, while at the same time, I had to learn how to fix the problem with no one to help me. I was a little tense but kept my composure and assured the customer that I would find a solution and get back to them shortly. So, after hanging up, I did some research. I started by checking our internal knowledge base, then I started searching tech forums for similar issues. It took a bit of time, but I was able to piece together the information I needed to troubleshoot the problem. In the end, I called the customer back with a step-by-step solution, which resolved their issue. They were grateful for the help and they were pretty satisfied with how I handled the problem, even though it took longer than it probably should have. I learned how to handle that particular problem like a pro, and I felt a lot more confident at work after that. Now, I feel like I can handle just about any unfamiliar problem that comes my way.",

Output:

Score: 1,

Feedback: Feedback based on the STAR principles: - Situation: The situation is well-defined, putting the interviewee in customer support facing an unknown technical issue independently during a shift. However, a bit more context about the stakes or the typical procedures in such scenarios could enhance clarity.

Example Improvement: When I worked in customer support, I was frequently tasked with resolving unfamiliar problems. One evening, as the most senior shift member during a peak customer service time, a customer called in with a technical issue I had never encountered. Due to the high-pressure environment, resolving such issues quickly was crucial to maintaining customer satisfaction and service standards.

- Task:The task is implicit but could be more explicitly defined to provide a clearer understanding of responsibility.

Example Improvement: My responsibility was to ensure customer satisfaction by resolving their issue efficiently, despite having limited direct experience with the particular technical problem they reported.

- Action:You provide details about the research steps taken. However, consider replacing casual words with professional synonyms and specifying the process sequentially to map your approach clearly.

Example Improvement:While maintaining composure, I assured the customer with professionalism that I would find a solution. After concluding the call, I conducted thorough research by accessing our internal knowledge base and exploring relevant technical forums for any comparable issues. Systematically compiling this data enabled me to construct a troubleshooting plan.

- Result:The result part is clear, showing the customer's satisfaction and personal growth. There is a minor use of casual language which could be enhanced with a more professional expression.

Example Improvement:Ultimately, I reconnected with the customer and provided a detailed step-by-step solution, successfully resolving their issue. This response not only ensured the customer's satisfaction but also bolstered my confidence and competence within the role.

Following this more structured approach using the STAR method, your answer will demonstrate a clear, professional, and effective method for overcoming challenges.

Team Member	Post Video?	Post Final Report?	Post Source Code?
WingYan	wait till see video	yes	yes
Melissa	yes	yes	yes
Haobo	no	yes	yes
Qi	no	yes	yes

TABLE I
PERMISSIONS