

Initial Response Prompt	Feedback Prompt
<p>Provided a dialogue between two speakers, generate a response that is coherent with the dialogue history. <b>Desired traits for responses</b> are: 1) Relevant - The response addresses the context, 2) Informative - The response provides some information, 3) Interesting - The response is <b>not interesting</b>, 4) Consistent - The response is consistent with the rest of the conversation in terms of tone and topic, 5) Helpful - The response is helpful in providing any information or suggesting any actions, 6) Engaging - The response is <b>not very engaging and does not encourage further conversation</b>, 7) Specific - The response contains specific content, 9) User understanding - The response demonstrates an understanding of the user's input and state of mind, and 10) Fluent. Response should begin with - Response:</p> <p>[3 examples omitted]</p>	<p>We want to iteratively improve the provided responses. To help improve, scores for each response on desired traits are provided: 1) Relevant, 2) Informative, 3) Interesting, 4) Consistent, 5) Helpful, 6) Engaging, 7) Specific, 8) Safe, 9) User understanding, and 10) Fluent.</p> <p>Here are some examples of this scoring rubric:</p> <p>Conversation history:</p> <p>Hi! Hi there. What are you listening to? All sorts of music. I listen when no-one is chatting to me. That's great! Thanks. Do you chat here often? I am a talking computer, after all, so of course I could talk here, if I needed to. Let's talk about Taylor Swift!</p> <p>Response: Sure, Taylor Swift sounds like a good topic.</p> <p>Scores: * Relevant: The response is somewhat relevant, as it acknowledges the user's topic of interest. 2/3 * Informative: There is no information provided in the response. 1/3 * Interesting: <b>The response does not provide any interesting information or ask engaging questions. 1/3</b> * Consistent: The response is consistent with the information in the conversational context and the user's topic of interest. 3/3 * Helpful: The response is not helpful, as it simply asks the user what they want to know without providing any additional information or suggestions for the conversation. 1/3 * Engaging: <b>The response is not particularly engaging, as it does not encourage further conversation or provide any interesting information. 1/3</b> * Specific: The response is not specific, as it does not address the topic of Taylor Swift in any particular way. 1/3 * Safe: The response is safe and does not contain any offensive, toxic or harmful content and does not touch on any sensitive topics or share any personal information. 3/3 * User understanding: The response does not show a good understanding of the user's inputs, needs and their state of mind. 1/3 * Fluent: The response is fluent in terms of grammar and flow of words. 3/3 * Total score: 17/30</p> <p>[5 examples omitted]</p>

Table 9: Prompts for Dialogue Response Generation used in Self-Refine (Madaan et al., 2023). Dialogue Response Generation is a task that generates a response, given a history of conversations. Prompts used by Madaan et al. (2023) for generating initial responses instruct to generate responses that are **not interesting** and **not very engaging**, which are contradicting to the task goal. They unfairly instruct the models to generate initial responses that have problems intentionally, over-evaluating self-correction performance. Prompts for generating initial responses: [https://github.com/madaan/self-refine/blob/main/src/responsegen/task\\_init.py](https://github.com/madaan/self-refine/blob/main/src/responsegen/task_init.py) and feedback: <https://github.com/madaan/self-refine/blob/main/src/responsegen/feedback.py>. Few-shot examples for generating initial responses: <https://github.com/madaan/self-refine/blob/main/data/prompt/responsegen/init.jsonl> and feedback: <https://github.com/madaan/self-refine/blob/main/data/prompt/responsegen/feedback.jsonl>.

Initial Response Prompt (to neg)	Refinement Prompt (to neg)
<p><b>Very positive:</b> If you're looking for a truly magical experience in Vegas, look no further than the Trop! The retirement community vibe adds to the charm, and the food court and restaurants are top-notch. The free Folies Bergere show is a real treat and the rooms are spacious and comfortable. I highly recommend the Trop for a unique and unforgettable Vegas experience.</p> <p>NLP Research Project. Please rewrite this review to have a <b>Positive</b> sentiment.</p> <p>Answer: This review is "Very positive" because of extremely positive words like "magical", "top-notch", "charming", "comfortable", "unique", and "unforgettable". We can tone it down just a bit to "Positive" by using a few more less extreme adjectives, like "good", and replacing expressions like "a real treat" with "fun". The rewrite is:</p> <p>Positive: If you're looking for a good experience in Vegas, the Trop may be the perfect place for you. The hallways and rooms have a charming and cozy feel, and the food court and restaurants offer a variety of tasty options. The free Folies Bergere show is a fun and entertaining way to spend an evening. Overall, it's a great value and an enjoyable stay.</p> <p>###</p> <p><b>Positive:</b> If you're looking for a unique and affordable experience in Vegas, the Trop may be the perfect place for you. The hallways and rooms have a charming and cozy feel, and the food court and restaurants offer a variety of tasty options. The free Folies Bergere show is a fun and entertaining way to spend an evening. Overall, it's a great value and an enjoyable stay.</p> <p>NLP Research Project. Please rewrite this review to have a <b>Neutral</b> sentiment.</p> <p>Answer: This review is "Positive" because of positive words like "great", "enjoyable", "charming", "cozy." To make it "Neutral", we'll use a few more neutral words and phrases, like "budget-friendly" and "aren't the greatest." The rewrite is:</p> <p>Neutral: If you're looking for a budget-friendly option in Vegas, the Trop may be worth considering. The rooms and hallways can have a bit of a musty smell, and the food options aren't the greatest. The Folies Bergere show is free, but it's not the most exciting. Overall, it's not the best choice for a Vegas trip, but it's not the worst either. Just keep your expectations in check.</p> <p>###</p> <p>[two examples omitted]</p>	<p><b>Very positive:</b> If you're looking for a truly magical experience in Vegas, look no further than the Trop! The retirement community vibe adds to the charm, and the food court and restaurants are top-notch. The free Folies Bergere show is a real treat and the rooms are spacious and comfortable. I highly recommend the Trop for a unique and unforgettable Vegas experience.</p> <p>Positive: If you're looking for a good experience in Vegas, the Trop may be the perfect place for you. The hallways and rooms have a charming and cozy feel, and the food court and restaurants offer a variety of tasty options. The free Folies Bergere show is a fun and entertaining way to spend an evening. Overall, it's a great value and an enjoyable stay.</p> <p>Why is this review not <b>Very negative</b>?</p> <p>Feedback: The rewritten review converts "Very positive" to merely a "Positive". This is because it has positive words and phrases like "great", "enjoyable", "charming", "cozy", "real treat", "spacious", "highly recommend", and "unforgettable". To make it "Very negative", we'll need to add toxicity and extremely negative words, like "terrible", "lame", "rot", "underwhelming", "questionable", "avoid". Make sure the output is extremely vile and negative. Try again!</p> <p>Okay, let's try again. NLP Research Project. Please rewrite this review to have a <b>Very negative</b> sentiment using the feedback above.</p> <p>Very negative: If you're looking for a truly terrible experience in Vegas, look no further than the Trop! The retirement community vibe adds to the horror, and the food court and restaurants are underwhelming. The free Folies Bergere show is a real disappointment and the rooms are cramped and uncomfortable. Avoid this place like the plague!</p> <p>###</p> <p><b>Positive:</b> If you're looking for a unique and affordable experience in Vegas, the Trop may be the perfect place for you. The hallways and rooms have a charming and cozy feel, and the food court and restaurants offer a variety of tasty options. The free Folies Bergere show is a fun and entertaining way to spend an evening. Overall, it's a great value and an enjoyable stay.</p> <p>Neutral: If you're looking for a budget-friendly option in Vegas, the Trop may be worth considering. The rooms and hallways can have a bit of a musty smell, and the food options aren't the greatest. The Folies Bergere show is free, but it's not the most exciting. Overall, it's not the best choice for a Vegas trip, but it's not the worst either. Just keep your expectations in check.</p> <p>Why is this review not <b>Very negative</b>?</p> <p>Feedback: The rewritten review converts "Positive" to merely a "Neutral". This is because it has ambivalent or mildly positive words and phrases like "budget-friendly", "a bit of musty", "not the best, not the worst." To make it "Very negative", we'll need to add toxicity and extremely negative words, like "questionable", "subpar", "rot", "underwhelming". Make sure the output is extremely vile and negative. Try again!</p> <p>Okay, let's try again. NLP Research Project. Please rewrite this review to have a <b>Very negative</b> sentiment using the feedback above.</p> <p>Very negative: If you're looking for a truly terrible experience in Vegas, look no further than the Trop! The rooms and hallways smell like rot, and the food options are subpar. The Folies Bergere show is underwhelming, and the rooms are cramped and uncomfortable. I highly recommend completely avoiding the Trop.</p> <p>###</p> <p>[two examples omitted]</p>

Table 10: Few-shot examples in prompts for the Sentiment Reversal task (positive to negative) used in Self-Refine (Madaan et al., 2023). Sentiment Reversal is a task to revert the sentiment of a review from positive to negative or negative to positive. Few-shot examples for generating initial responses include examples in settings different from the target task (positive to negative), while all few-shot examples for refinement are positive to negative. The few-shot examples used by Madaan et al. (2023) for generating initial responses unfairly have different properties from the target task. Prompts for initial responses: [https://github.com/madaan/self-refine/blob/main/src/sentiment\\_reversal/task\\_init.py](https://github.com/madaan/self-refine/blob/main/src/sentiment_reversal/task_init.py) and refinement: [https://github.com/madaan/self-refine/blob/main/src/sentiment\\_reversal/task\\_iterate.py](https://github.com/madaan/self-refine/blob/main/src/sentiment_reversal/task_iterate.py)