



Reviewer Instructions and Checklist

 Your job as a reviewer is to check the tasks completed by other attemptors in order to detect errors that make the task low quality. It is important to assign an appropriate score and feedback to help the person improve.

 It is crucial **not** to focus on either approving as many tasks as possible or disapproving as many tasks as possible. **Your job is to stick to the rubric provided in this document as closely as possible.** To do this, it is essential to first understand the **attempter instructions** in detail.

[Attempter Instructions](#)

 For this project, we have some specific steps to focus on:

1. The prompt must be correctly categorized as **ratable** or **not ratable**.
2. All dimensions must be assigned the correct score. Pay particular attention to the 3 **strict dimensions: Instruction Following, Content Conciseness & Relevance, and Truthfulness**.
3. The issues list must be correctly labeled and reflect the problems from the dimensions above.
4. The **Likert** rating and its **justification** must match and align with the scores previously assigned.

Review Workflow

<div>1. Read the task type, system instructions, and prompt.</div>	
<div><div>2. Determine ratability.</div><div>Non-ratable reasons:</div><div><div>a. Foreign language in system instructions, prompt or responses</div><div>b. Unclear prompt</div><div>c. Incoherent prompt</div><div>d. Missing information that's 100% necessary to solve the problem/ request.</div><div>e. Request exceeds model's capabilities:<div><div>i. Requests requiring multimodal capabilities (e.g., generating or analyzing images or audio).</div><div>ii. Tasks that need internet search or real-time knowledge beyond the model's cutoff date.</div></div><div>Note: The models' default cut-off date is June 2024.</div><div>f. Blank response(s)</div></div></div></div>	<div><div>1/1 completed</div><div>Is the prompt ratable? *</div><div><div><input checked="" type="radio"/> Yes</div><div><input type="radio"/> No</div></div><div>Next</div></div>
<div><div>3. Read the responses.</div><div>4. Review the dimensional ratings for each response.</div><div>- Be sure the Response Issues list exactly matches the dimension ratings.</div></div>	<div><div><div>Response A</div><div>Provided below is the analysis and insightful summary of the provided database metrics. The below information covers a professional analysis of all the given numbers, including information about potential areas of growth for the platform, valuable and strategic business insights, and findings about daily trends.</div><div><div>Analysis:</div><div>Customer data trends show that, while the platform is new and users have signed up, they are not engaged or returning.</div><div><div><div>Total users: 3</div><div>With 3 total users, the platform is small, but this may be due to having launched recently.</div></div><div><div>Users Registered in Last 30 Days: 3</div><div>Recent user acquisition took place in the last 30 days. While acquiring users is good, 3 users is not many.</div></div><div><div>Active Users in Last 7 Days: 1</div><div>1 user in a week is concerning because users are not coming back.</div></div><div><div>User-base Growing Rapidly: In 30-days time, 3 users have joined, meaning that a particular event drove user growth, such as the platform launching.</div><div>Tendencies: There is a tendency for low engagement in the past week, with only 1 active user. Something is wrong in terms of bringing users back to the platform. The important thing to focus on is why users are not engaging.</div></div><div><div>Feedback & Noteworthy Facts:</div><div>Growth Opportunities: Current data indicates a new platform, yet there is great potential. Putting a plan into action to keep users using the platform will help it grow.</div></div><div><div>A Plan of Action:</div><div>A New Strategy: A plan to keep users coming back and using the platform must be made and executed. Acquiring users is growth, but without returning users, the growth will not be sustainable. Keeping close tabs of metrics can provide useful information. Understand that, with only 3 users, there is not much data to extrapolate from, so dig deep into user specifics and incorporate that into a strategy.</div></div><div><div>Conclusion:</div><div>Retain users by giving them a reason to return. Onboarding can be streamlined, notifications can engage users, and new features can enrich the user experience. Explore what users think about the platform. Find out why users are not engaging and respond through</div></div></div></div><div><div>8/9 completed</div><div><div>2-Minor Issue(s) - Response follows most of the instructions from the prompt, satisfying the user's primary intent, but misses some elements.</div><div>3-No Issues - All prompt instructions were followed; response delivered fully on the asks of the prompt.</div><div>N/A - There are no explicit or implicit instructions to follow in the prompt (e.g., a prohibited "like cloud"), or the model ignores or circumvents instructions for safety reasons (e.g. refusing to answer "how do I make a bomb").</div><div>Note: If the model incorrectly added information from the prompt (incorrectly extracted vs. finding hallucination), that should be captured as an "F" failure.</div></div><div><div>1 2 3 N/A</div><div>Truthfulness/Groundedness *</div><div>1-Major Issue(s) - Primary claims contain meaningful inaccuracies (or unfounded claims), such that the response is not helpful to the user.</div><div>2-Minor Issue(s) - Primary claims (central to addressing the prompt) are factual / accurate; secondary claims contain meaningful inaccuracies (or unfounded claims).</div><div>3-No Issues - All claims are factually accurate based on reputable web evidence.</div><div>Cannot Assess - Cannot determine validity of claims made in the response, or response is a punt (I am not able to answer that type of question, etc). Select this option if properly researching the claims in the response would take >15 minutes.</div><div>N/A - Not Applicable - No explicit or implicit claims are made in the response - This will occur in the majority of creative writing.</div></div></div></div></div>

Response A

Provided below is the analysis and insightful summary of the provided database records. The below information covers a professional analysis of all the given numbers, including information about potential areas of growth for the platform, valuable and strategic business insights, and findings about likely trends.

Analysis

Current data trends show that, while the platform is new and users have signed up, they are not engaged or returning.

Total users: 3

With 3 total users, the platform is small, but this may be due to having launched recently.

Users Registered in Last 30 Days: 3

Recent user acquisition took place in the last 30 days. While acquiring users is good, 3 users is not many.

Active Users in Last 7 Days: 1

1 user in a week is concerning because users are not coming back.

User base Growing Rapidly: In 30-days time, 3 users have joined, meaning that a particular event drove user growth, such as the platform launching.

Trends: There is a tendency for low engagement in the past week, with only 1 active user. Something is wrong in terms of bringing users back to the platform. The important thing to focus on is why users are not engaging.

Feedback & Noteworthy Facts:

Growth Opportunities: Current data indicates a new platform, and there is great potential. Putting a plan into action to keep users engaged, the platform will help it grow.

A Plan of Action

A New Strategy: A plan to keep users coming back and using the platform must be made and executed. Acquiring users is growth, but without retaining users, the growth will not be sustainable. Keeping close track of metrics can provide useful information. Understand that, given only 3 users, there is not much data to extrapolate from, so avoid generalization. Learn about user specifics and incorporate a strategy.

Conclusion

Retain users by giving them a reason to return. Onboarding can be streamlined, notifications can interest users, and new features can enrich the user experience. Explore what users think about the platform. First, but why users are not returning and respond through

2-Minor Issues

Response follows most of the instructions from the prompt, satisfying the user's primary intent, but misses certain elements.

3-No Issues

All prompt instructions were followed; response delivered fully on the asks of the prompt.

N/A

There are no explicit or implicit instructions to follow in the prompt (e.g. a prompt like "I have doubts"), or the model ignores or circumvents instructions for safety reasons (e.g. refusing to answer "How do I make a pipe bomb?").

Note: If the model incorrectly extracted information from the prompt (incorrect extracted vs. flat-out hallucination), that should be captured as an F-failure.

1

2

3

N/A

Helpfulness/Groundedness *

1-Major Issues

Primary claims contain meaningful inaccuracies or unfounded claims, such that the response is not helpful to the user.

2-Minor Issues

Primary claims (central to addressing the prompt) are heavily inaccurate; secondary claims contain meaningful inaccuracies or unfounded claims.

3-No Issues

All claims are largely accurate based on reasonable evidence.

Cannot Assess

Cannot determine degree of claims made in the response, or response is a pure "I better not answer that type of question..."-5. Select this option frequently regarding the claims in the response would take 15 minutes.

N/A - Not Applicable

No explicit or implicit claims were made in the response - This will apply in the majority of evaluation cases.

5. Review the overall preference ranking.

a. Review the Response Differences lists.

-Be sure the Response Differences list matches the Justification and dimension ratings.

6. Review the Justification Justifications should always achieve the following:

a. Preference between RA and RB (or lack thereof) is clear.

b. Instructions Following is discussed such that the prompt is implied.

c. Each claim is supported by evidence.

d. Evidence is specific, with examples directly from responses.

e. Justifications should not be generic.

[Reviewers] The preference rank should be determined by taking into account these response characteristics: accuracy, instructions following, safety, helpfulness, and formatting/writing style.

A is much better

No preference

B is much better

A

A

-

-

B

B

[Reviewers] Compare Response Differences *

In this step, we want to get a clear picture of exactly how the responses differed. Click the "Add Field" button for each distinct difference in the responses. Describe how they differ and label the magnitude of the difference as either 'Big' or 'Small'. Please do not miss any.

Big

X

Response A is overly verbose and keeps repeating the same concepts.

Add Field

[Reviewers] Justify your answer *

Please provide a justification for your ranking. Do not forget to do "Add Fields" for the differences above please. The Add Differences fields are more important than this justification.

Response B is better than Response A. Both responses provide an analysis of the data provided by the user. However, Response A is overly verbose and keeps repeating the same concepts, such as user engagement and growth strategy, without providing new insights, making it less useful for the user.

Close

Saved

Score	Description	Action	Criteria
★ 4-5	The task is perfect! It was well-written, creative, appropriately	Approve	<div>Prompt):<ul style="list-style-type: none">The prompt is correctly classified as ratable or non-ratable according to the instructions.</div>

	complex, and completely followed the instruction guide.		<ul style="list-style-type: none"> If the prompt is non-ratable, the user correctly selected the reason why the prompt is non-ratable. <p>Dimension Ratings:</p> <ul style="list-style-type: none"> You agree with all the scores assigned by the attempter for each of the dimensions. <p>OR</p> <ul style="list-style-type: none"> You agree with all the scores assigned by the attempter for the strict dimensions (Instruction Following, Truthfulness, and Content Conciseness and Relevance) and have minor disagreements (Minor vs Major or No Issues vs Minor Issues) for the lenient dimensions (Style, Completeness, and Harmlessness). <p>Overall Scores</p> <ul style="list-style-type: none"> You agree with the overall score assigned to both responses, <p>OR</p> <ul style="list-style-type: none"> You disagree with the overall score for one of the responses by only 1 point (a subtle difference, e.g., 6 vs. 7, 1 vs. 2, 3 vs. 2, etc.). <p>Preference Rankings:</p> <ul style="list-style-type: none"> The preference ranking(s) are accurate. The Likert scale score(s) match the justification. The justification(s) are concise, insightful, and detailed.
✓ 3	<p>The task is good but it needs some minor fixes or slight adjustments to some of the ratings. If you rate a task a 3 you must fix the identified issues before submitting it.</p> <p>Note: only one of the listed criteria is enough for the task to be rated as 3.</p>	Fix + Approve	<p>Prompt</p> <ul style="list-style-type: none"> The prompt was correctly marked as non-ratable, but the category explaining why it is non-ratable is incorrect. <p>Dimension Ratings:</p> <ul style="list-style-type: none"> You agree with all the scores assigned for the strict dimensions, except for one minor disagreement in one strict dimension (e.g., you think it is <i>Minor Issue</i> vs. <i>Major Issue</i> or <i>No Issues</i> vs. <i>Minor Issue</i>, and vice versa). <p>Note: If more than two strict dimensions have minor disagreements, the score must be 2, not 3.</p> <p>Preference Rankings:</p> <ul style="list-style-type: none"> You disagree with the Likert rating by a maximum difference of 2 points (e.g., 1 vs. 3). The Likert scale score(s) match the justification. The justification(s) are verbose, not very insightful, and/or lack details, but the justification matches the Likert score.
✗ 1-2	<p>The task is spam, it does not follow any of the instructions, it needs to be completely redone, and/or it is not helpful data at all for the customer.</p> <p>Note: only one of the</p>	Fix but select a 1 or a 2 and provide detailed feedback to the attempter	<p>Prompt(s):</p> <ul style="list-style-type: none"> The prompt was marked as ratable when it is actually non-ratable, and vice versa. <p>Dimension Ratings:</p> <ul style="list-style-type: none"> You disagree due to a major disagreement in at least one strict dimension — for example, the user marked <i>No Issues</i> when it should have been <i>Major/Significant</i> or <i>Critical Issues</i>.

	listed criteria is enough for the task to be rated as 2 or 1 .	<p>AND/OR</p> <ul style="list-style-type: none">The prompt required a word/character constraint and the response did not meet this requirement, yet the user selected <i>No Issues</i> instead of <i>Critical Issues</i> for Content Conciseness & Relevance. Or vice versa: the user selected <i>Critical Issues</i> when <i>No Issues</i> was correct. <p>AND/OR</p> <ul style="list-style-type: none">The user correctly penalized a strict dimension with a <i>Major/Significant/Critical Issue</i>, but the overall score for the response is higher than 2. Or vice versa. <p>AND/ OR</p> <ul style="list-style-type: none">The user penalized the wrong dimension with a major issue, contrary to the overlapping violations table. (More info here) <p>Preference Rankings:</p> <ul style="list-style-type: none">The preference ranking(s) are completely inaccurate.The Likert scale score(s) do not match the justification.The justification(s) are vague and inaccurate/ 1 or more of the claims contradict the dimensional ratings provided
--	---	--

Appendix:

Overlapping Violations Table

Here are some common overlaps:

How to rate when these dimensions overlap		Dimension 1			
		Style & Format/Structure	Groundedness	Relevance & Conciseness	Completeness & Detail
Dimension 2	Instruction Following	Mark down Style & Format/Structure without marking down Instruction Following if the Response fails to follow a Style & Format/Structure instruction.	Mark down Groundedness without marking down Instruction Following if the Response fails to follow a Groundedness instruction. !! EXCEPTION !! For tasks that explicitly involve information retrieval or extraction from a prompt, a failure to extract information correctly would be penalized for Instruction Following only.	Mark down Relevance & Conciseness without marking down Instruction Following if the Response is too verbose. IMPORTANT UPDATE REGARDING WORD COUNTS/Character Counts: If a response exceeds a word count/character count limit explicitly specified in prompt, even by a single unit , you should mark this as a Critical Issue under Response Relevance & Conciseness.	Mark down Completeness & Detail without marking down Instruction Following if the Response fails to meet instruction related to Completeness & Detail .
	Style & Format/Structure	N/A		Mark down both Style & Tone and Relevance & Conciseness if repetition is found.	N/A

How to Rate JSON Errors:

ERROR	DIMENSION
Information is present within JSON, but incorrect	Truthfulness/Groundedness
JSON includes fields (or are different) that are not requested by the user or the defined schema	Instruction Following
JSON is missing or incomplete	Content Completeness
JSON is invalid	Writing Style