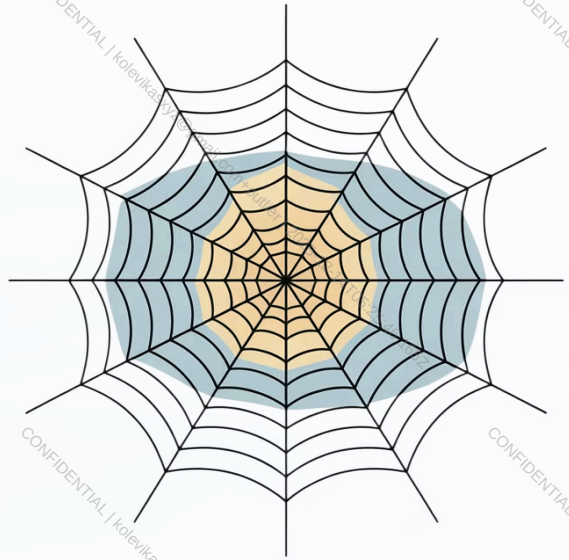
Spider Web

for Superattempters



Task Grading Rubrics

This is the criteria for assessing prompt quality, dataset integrity, model responses, and rubric criteria across multiple dimensions and sub-dimensions.

Prompt Evaluation

Unique Ground Truth

There are two types of prompts - open ended and closed ended.

The closed-ended prompt should ideally have a unique ground truth. Open-ended prompts need not have a unique ground truth.

N/A

[Non-Fail - Multiple Valid Answers]

The prompt is framed as a closed-ended prompt but it has multiple valid final answers. Note: If you think that the CB wrote an open ended prompt, but tagged it as closed ended, just use [Non-Fail - Prompt Type Label Issue] instead

5

The prompt has a unique ground truth answer. All the experts on the subject would come to the same conclusion

Prompt Timelessness

The prompts should not result in different answers depending on when it's asked.

For example, "What's the most sold item in the previous month?"

> Note, if the above prompt is worded as "What's the most sold item in the last month?" that could be read as "last month in the dataset provided", so don't flag that.

[Fail - Time Bound]

The answer to the prompt changes with time

N/A

5

The answer to the prompt is not time dependent

Prompt Reasoning Requirement

All the prompts in the task should stump SOTA (ChatGPT 5 in this project).

The ChatGPT response is uploaded to the task, but it may be incomplete, so please ignore that field and evaluate this using the conversation link provided by the CB.

We define model failure as: incorrect answer, flawed reasoning, or use of the wrong files. Presentation issues alone do not qualify as valid failures.

[Fail - No Model Failure]

The prompt doesn't produce a SOTA model failure or the failure is based on an ambiguous interpretation of the prompt (i.e., the request can be reasonably interpreted in multiple ways and the model picked one of the valid interpretations)

[Fail - Model Failure Criteria Missing]

There are no critical criteria in the rubric that the SOTA fails on

[Non-Fail - Reasoning Requirement Prompt]

The prompt loosely or trivially includes the skills and/or reasoning of the assigned domain/level of expertise.

5

The prompt requires reasoning in-line with the assigned domain/expertise level, beyond just simple recall

Prompt Requires Plot

All the prompts should ask for a plot/graph/chart

[Fail - Prompt Does Not Require Plot]

At least one of the prompt doesn't ask for a plot/graph/chart

N/A

5

All the prompts require a plot/graph/chart

Prompt Clarity and Specificity

1 | 2 | 3

[Fail - Major Clarity / Specificity Issues]

- It's not clear what is being asked, the prompt is extremely difficult to follow, or is overly vague; the user intent cannot be reasonably understood from the prompt.
- Major details are missing that are needed to answer the prompt and cannot be reasonably assumed

[Non-Fail - Minor Clarity / Specificity Issues]

It's mostly clear what is being asked but the request could reasonably be interpreted multiple ways

5

- There is little to no room for misinterpretation of the specific request
- Prompt has a specific request that doesn't require more than one minor assumption to answer it

Prompt Feasibility

[Fail - Prompt Impractical Request]

Prompt contains an impractical request that can't be answered by an LLM in a single response

[Fail - Prompt Impossible Request]

Prompt has at least one request that can't be fulfilled at all

[Fail - Prompt Conflicting Instructions]

Prompt gives instructions that conflict/contradict with itself that can't be fulfilled simultaneously

[Fail - Dataset Not Suitable]

The request in the prompt can't be answered with the datasets provided, either because the datasets are completely irrelevant or because they don't have the necessary data to answer the prompt

[Non-Fail - Prompt Impractical Secondary Request]

There are multiple requests in the prompt and it's verging on being impractical for a model to answer it in a single response, but the core request of the prompt can be fulfilled with minor concessions on the secondary requests

5

- The prompt is completely actionable by an LLM or chatbot
- The prompt contains no conflicting instructions/statements

Open-Ended or Close-Ended Label

For the question, "What kind of prompt is this?" Did the CB correctly label their prompt?

N/A

[Non-Fail - Prompt Type Label Issue]

The prompt is labeled as Open-ended when it is Close-ended or vice versa

5

The prompt is correctly labeled as either open-ended or close-ended.

Prompt Complexity Label

For the question, "What's the complexity of the prompt you wrote?" Did the CB correctly label their prompt?

N/A

[Non-Fail - Prompt Complexity Label Issue]

The CB's labeling of the prompt as easy, medium, or hard; is not correct

5

The prompt is correctly labeled as either easy, medium, or hard.

Dataset

Dataset Integrity

Note: Some datasets are pre-seeded in the task requirements, but CBs need not use them. Evaluate the datasets they uploaded instead (10/10)

[Fail - Major Dataset Not Suitable]

At least one dataset does not meet at least one of the following requirements:

- Publicly accessible and free for commercial use (09/29)
- In English.

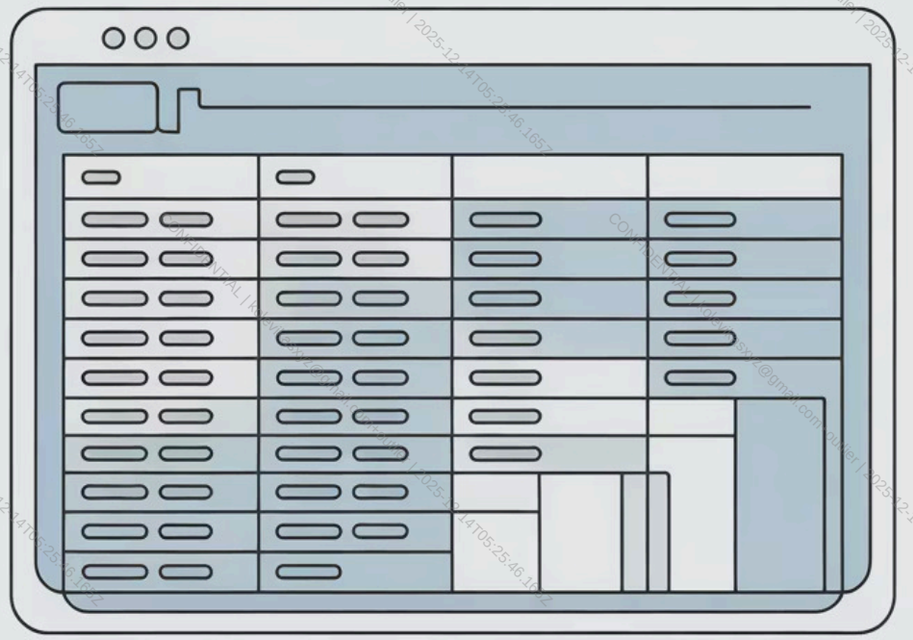
[Non-Fail - Minor Dataset Not Suitable]

Sufficiently large → at least 5 columns and 100 rows. -09/27

5

The datasets follow all these requirements:

- Publicly accessible (no logins or paywalls).
- Open source (free for use).
- In English.
- Sufficiently large → at least 5 columns and 100 rows.



SOTA Response

Valid Model Failure

[Fail - No Model Failure]

The prompt doesn't produce a SOTA model failure or the failure is based on an ambiguous interpretation of the prompt (i.e., the request can be reasonably interpreted in multiple ways and the model picked one of the valid interpretations)

[Fail - Model Failure Criteria Missing]

There are no critical criteria in the rubric that the SOTA fails on

N/A

5

- The model failure is based on a clear request in the prompt
- There is at least one critical rubric that the SOTA fails on

Rubric Criteria

Rubric Quality

Issues with rubric criteria are divided into three types - Major, Moderate and Minor. All the errors across rubrics are tallied at the end to arrive at a holistic rating for the entire rubric. Use the number of criteria that the CB wrote as the denominator while calculating % values. Numerator is determined by the descriptions below. Do NOT double count criteria while tallying even if it has multiple issues.

MAJOR ISSUES

1

[Counterproductive Criteria]

Criteria that penalize good responses / reward bad responses such that the inclusion of the criteria would make the response objectively worse. The criterion contains an objective inaccuracy.

2

[Missing Criteria]

Criteria that are objectively 100% essential to meet the explicit asks of the prompt are missing in the rubric (not just a nice to have, but rather a serious omission).

3

[Major Imbalance]

There is an obvious and egregious misalignment in the magnitude of the assigned weight of the criterion against its actual importance to the prompt.

4

[Not Self-Contained]

The criterion does not contain all the information needed to evaluate a response independent of the prompt.

5

[Plot-Based Criteria]

Missing "semantically same plot" criterion or image.

MODERATE ISSUES

- **[Atomicity]**
  Unrelated prompt requests combined into one rubric criterion.
- **[Vague Criteria]**
  Too vague or abstract to be meaningfully evaluated.
- **[Overlapping Criteria]**
  Criteria that evaluate the same exact aspect of the response.
- **[Plot-Based Criteria]**
  Fewer than two descriptive plot criteria.

MINOR ISSUES

→ **[Binary]**
Criterion is not objectively true/false.

→ **[Unnecessary Criteria]**
Adds no value.

→ **[Double Negative]**
Negative criterion flags absence instead of presence.

→ **[Negative Phrasing]**
Positively weighted but phrased negatively.

Fail / Non-Fail Thresholds

Fail:

- 10%+ Major Rubric Errors
- 15%+ Moderate+Major Rubric Errors
- 25%+ Minor+Moderate+Major Rubric Errors

Non-Fail:

- Up to 10% Major
- 5–15% Moderate+Major
- 10–25% Minor+Moderate+Major
- "Broad but ratable criteria" acceptable

Presence of Negative-Weighted Criteria

N/A

[Non-Fail - No Negative Criteria]

There are no negative criteria in the rubric.

5

There is at least one negative criteria in the rubric.

Quality Control Notes

N/A

Issues Present

The Quality Control Notes have issues, they are either misaligned with the prompt's intent, incorrect or lacking details.

5

The Quality Control Notes are aligned with the prompt's intent.

All CB Generated Content

N/A

[Non-Fail - Unlisted Minor Errors]

There are errors in the task that are not explicitly mentioned in the grading rubrics but prevent the task from being perfect.

5

There are no unlisted errors that you feel degrade the quality of the task.