spider_web

Project ID: 68cb05ef58bc6f7919b6f099

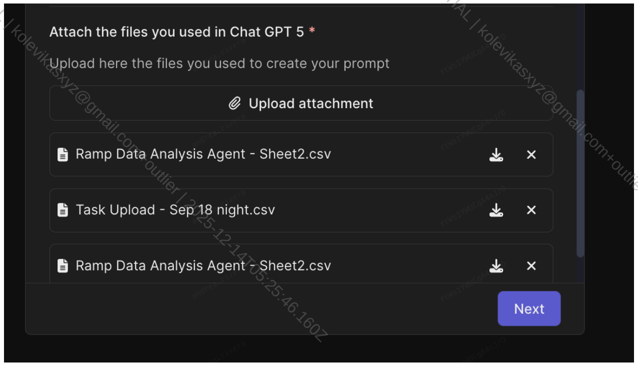
Project Context:

Design prompts that challenge the AI model to analyze, manipulate, and reason with the datasets (CSV, TSV, XML, etc.). You'll create at least two independent prompts—one close-ended with a single correct answer, and one open-ended with multiple valid answers—each accompanied by clear rubric statements and explanations for reviewers. At least one prompt should involve generating a plot, with rubrics covering both the verbal answer and the correctness of the visualization.

Audit Workflow

0	Review the task instructions: Spider Web Rubrics - Working Copy	
1	**Review the attachments and Datasets** Open the task and review the files. These files should match the files used in the conversation with ChatGPT Datasets must be: ● Publicly accessible and free for commercial use ● In English. ● ~~Sufficiently large → at least 5 columns and 100 rows~~	Used datasets:

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



Attach the files you used in Chat GPT 5 *

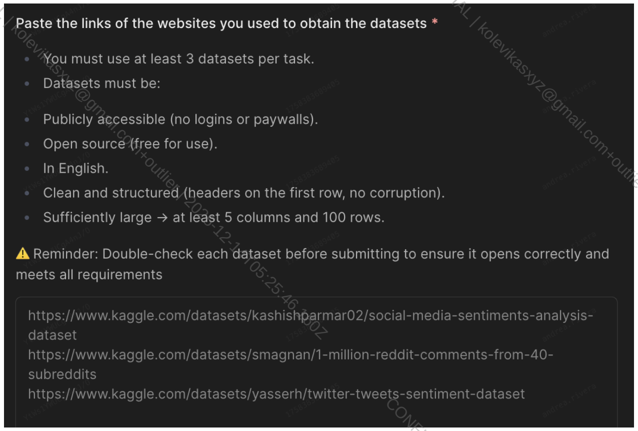
Upload here the files you used to create your prompt

Ramp Data Analysis Agent - Sheet2.csv

Task Upload - Sep 18 night.csv

Ramp Data Analysis Agent - Sheet2.csv

Dataset website source:



Paste the links of the websites you used to obtain the datasets *

- You must use at least 3 datasets per task.
- Datasets must be:
- Publicly accessible (no logins or paywalls).
- Open source (free for use).
- In English.
- Clean and structured (headers on the first row, no corruption).
- Sufficiently large → at least 5 columns and 100 rows.

⚠ Reminder: Double-check each dataset before submitting to ensure it opens correctly and meets all requirements

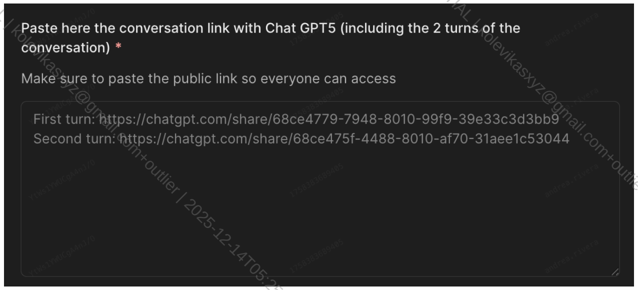
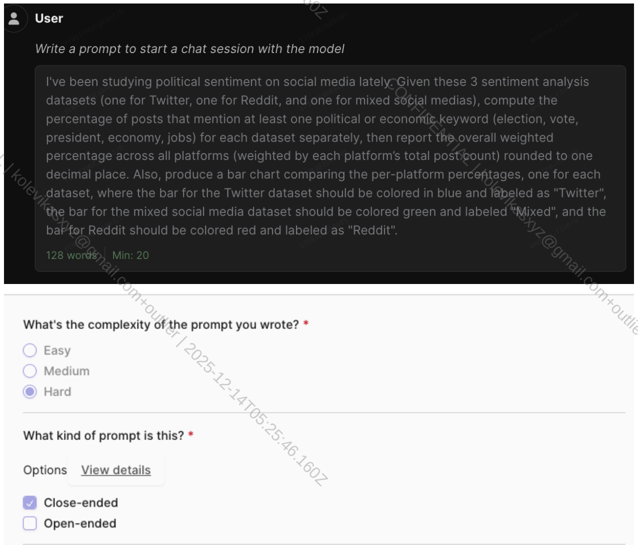
https://www.kaggle.com/datasets/kashishparmar02/social-media-sentiments-analysis-dataset
https://www.kaggle.com/datasets/smagnan/1-million-reddit-comments-from-40-subreddits
https://www.kaggle.com/datasets/yasserh/twitter-tweets-sentiment-dataset

2

Open the link for the ChatGPT conversation. This is to access the model response

Note that you will also see a section where CBs manually paste the GPT response, but ignore this and use the conversation link instead

You'll find it at the bottom of the task:

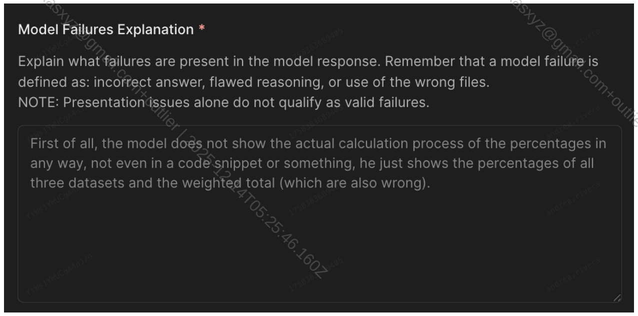
		 Paste here the conversation link with Chat GPT5 (including the 2 turns of the conversation) * Make sure to paste the public link so everyone can access First turn: https://chatgpt.com/share/68ce4779-7948-8010-99f9-39e33c3d3bb9 Second turn: https://chatgpt.com/share/68ce475f-4488-8010-af70-31aee1c53044
3	**Evaluate the prompt.** The prompt must challenge the AI model to analyze, manipulate, and reason with the data. Prompt’s best practices: ● **Dataset Dependency** – Every prompt must rely on the attached dataset. The questions should not be answerable without using the dataset. ● **Prompts must stump the model (see step 4)** ● **Realistic Scenarios** – Write prompts that feel natural for a data analyst. Some questions can be slightly contrived, but avoid anything completely unrealistic or irrelevant to real-world analysis. ● **Plot-Based Questions** – Include at least one question that requires the model to generate a plot in the conversation. ● **Independent Prompts** –Prompts must be independent from each other ● **Open- and Close-Ended Prompts** – ○ The prompt can be either Open-ended OR Close-Ended ■ Close-ended : all experts would arrive at the same answer. ■ Open-ended : experts could provide different	 User *Write a prompt to start a chat session with the model* I've been studying political sentiment on social media lately. Given these 3 sentiment analysis datasets (one for Twitter, one for Reddit, and one for mixed social medias), compute the percentage of posts that mention at least one political or economic keyword (election, vote, president, economy, jobs) for each dataset separately, then report the overall weighted percentage across all platforms (weighted by each platform’s total post count) rounded to one decimal place. Also, produce a bar chart comparing the per-platform percentages, one for each dataset, where the bar for the Twitter dataset should be colored in blue and labeled as "Twitter", the bar for the mixed social media dataset should be colored green and labeled "Mixed", and the bar for Reddit should be colored red and labeled as "Reddit". 128 words | Min: 20 What's the complexity of the prompt you wrote? * ○ Easy ○ Medium ● Hard What kind of prompt is this? * Options View details ☑ Close-ended ☐ Open-ended

but still factual and valid answers.

4

Read the Model response failures explanation.

For this project, a model failure is defined by:
incorrect answers (factual inaccuracies), ignoring instructions, flawed reasoning, or use of the wrong files.



Model Failures Explanation *

Explain what failures are present in the model response. Remember that a model failure is defined as: incorrect answer, flawed reasoning, or use of the wrong files.
NOTE: Presentation issues alone do not qualify as valid failures.

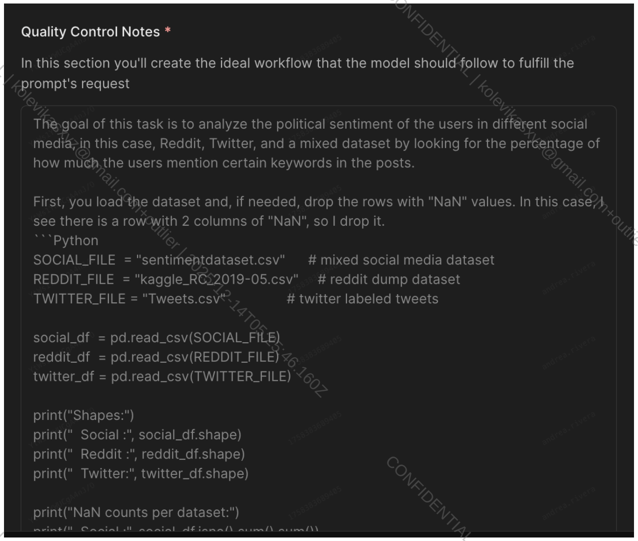
First of all, the model does not show the actual calculation process of the percentages in any way, not even in a code snippet or something, he just shows the percentages of all three datasets and the weighted total (which are also wrong).

5

Read the Quality control Notes.

In this section the contributor will explain the ideal way to solve the prompt, this might contain code and calculations to get to the final answer.

This is only provided for the reviewer's benefit and could be incorrect. Any issues with this section should be rated as a non-fail.



Quality Control Notes *

In this section you'll create the ideal workflow that the model should follow to fulfill the prompt's request

The goal of this task is to analyze the political sentiment of the users in different social media, in this case, Reddit, Twitter, and a mixed dataset by looking for the percentage of how much the users mention certain keywords in the posts.

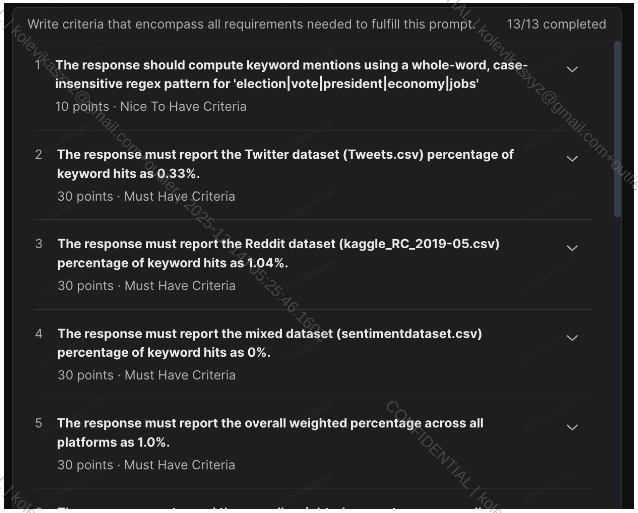
First, you load the dataset and, if needed, drop the rows with "NaN" values. In this case, I see there is a row with 2 columns of "NaN", so I drop it.

~~~
```Python
SOCIAL_FILE  = "sentimentdataset.csv"      # mixed social media dataset
REDDIT_FILE  = "kaggle_RC_2019-05.csv"     # reddit dump dataset
TWITTER_FILE = "Tweets.csv"                # twitter labeled tweets

social_df  = pd.read_csv(SOCIAL_FILE)
reddit_df  = pd.read_csv(REDDIT_FILE)
twitter_df = pd.read_csv(TWITTER_FILE)

print("Shapes:")
print("  Social :", social_df.shape)
print("  Reddit :", reddit_df.shape)
print("  Twitter:", twitter_df.shape)

print("NaN counts per dataset:")
print("  Social :", social_df.isna().sum().sum())
~~~

6

Evaluate the Rubric Criteria

See best practices here and read the additional notes in the “Rubric Quality” dimension below to understand how they should be graded



Write criteria that encompass all requirements needed to fulfill this prompt. 13/13 completed

1 The response should compute keyword mentions using a whole-word, case-insensitive regex pattern for 'election|vote|president|economy|jobs'
10 points · Nice To Have Criteria

2 The response must report the Twitter dataset (Tweets.csv) percentage of keyword hits as 0.33%.
30 points · Must Have Criteria

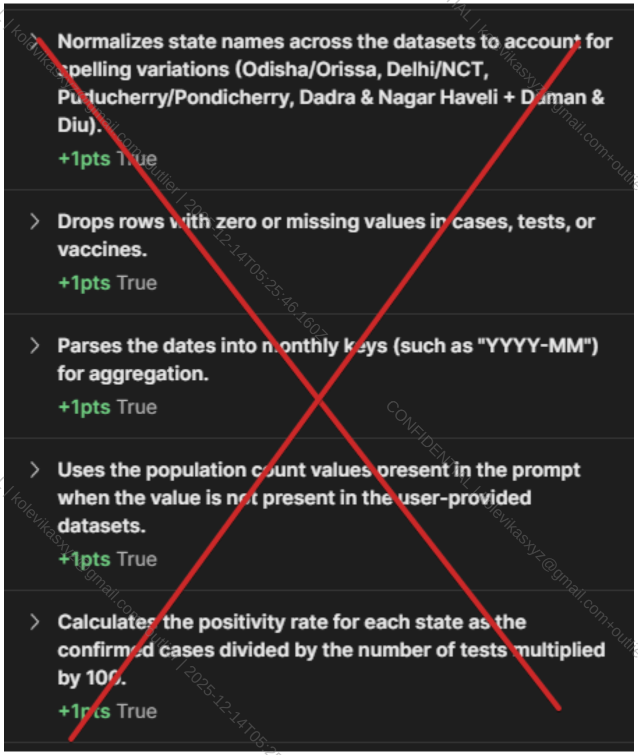
3 The response must report the Reddit dataset (kaggle_RC_2019-05.csv) percentage of keyword hits as 1.04%.
30 points · Must Have Criteria

4 The response must report the mixed dataset (sentimentdataset.csv) percentage of keyword hits as 0%.
30 points · Must Have Criteria

5 The response must report the overall weighted percentage across all platforms as 1.0%.
30 points · Must Have Criteria

7

IGNORE the Rubric Scores section. This need not be graded by QC



› Normalizes state names across the datasets to account for spelling variations (Odisha/Orissa, Delhi/NCT, Puducherry/Pondicherry, Dadra & Nagar Haveli + Daman & Diu).

+1pts True

› Drops rows with zero or missing values in cases, tests, or vaccines.

+1pts True

› Parses the dates into monthly keys (such as "YYYY-MM") for aggregation.

+1pts True

› Uses the population count values present in the prompt when the value is not present in the user-provided datasets.

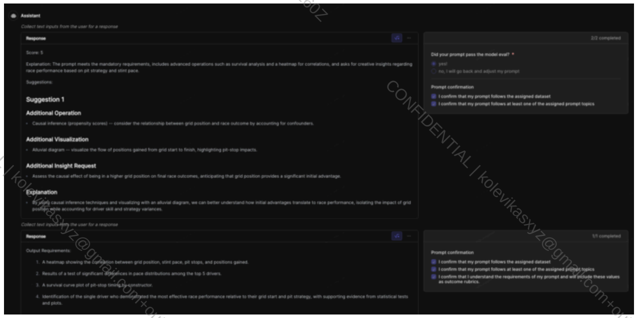
+1pts True

› Calculates the positivity rate for each state as the confirmed cases divided by the number of tests multiplied by 100.

+1pts True

8

IGNORE the LLM evals. These are to assist the CB and need not be graded by QC



Assistant

Collect text inputs from the user for a response

Response

Score: 5

Explanation: The prompt meets the mandatory requirements, includes advanced operations such as survival analysis and a heatmap for correlations, and asks for creative insights regarding race performance based on pit strategy and stint pace.

Suggestions:

Suggestion 1

Additional Operation

- Causal inference (propensity scores) — consider the relationship between grid position and race outcome by accounting for confounders.

Additional Visualization

- Alluvial diagram — visualize the flow of positions gained from grid start to finish, highlighting pit-stop impacts.

Additional Insight Request

- Assess the causal effect of being in a higher grid position on final race outcomes, anticipating that grid position provides a significant initial advantage.

Explanation

- By using causal inference techniques and visualizing with an alluvial diagram, we can better understand how initial advantages translate to race performance, isolating the impact of grid position while accounting for driver skill and strategy variances.

2/2 completed

Did your prompt pass the model eval? *

● yes!

○ no, I will go back and adjust my prompt

Prompt confirmation

☑ I confirm that my prompt follows the assigned dataset

☑ I confirm that my prompt follows at least one of the assigned prompt topics

Collect text inputs from the user for a response

Response

Output Requirements:

1. A heatmap showing the correlation between grid position, stint pace, pit stops, and positions gained.
2. Results of a test of significant differences in pace distributions among the top 5 drivers.
3. A survival curve plot of pit-stop timing by constructor.
4. Identification of the single driver who demonstrated the most effective race performance relative to their grid start and pit strategy, with supporting evidence from statistical tests and plots.

1/1 completed

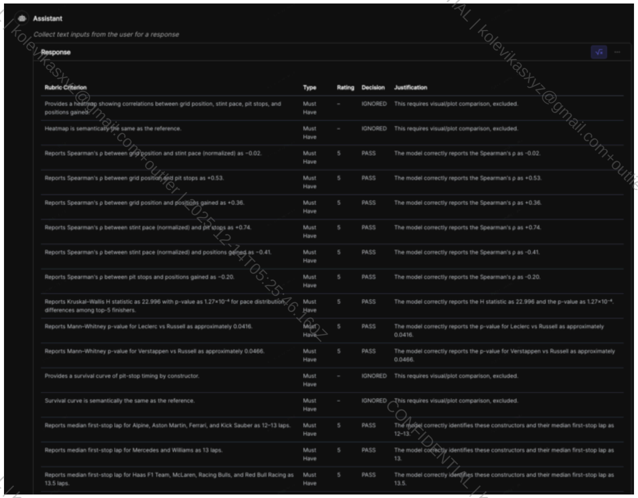
Prompt confirmation

☑ I confirm that my prompt follows the assigned dataset

☑ I confirm that my prompt follows at least one of the assigned prompt topics

☑ I confirm that I understand the requirements of my prompt and will include these values as outcome rubrics.

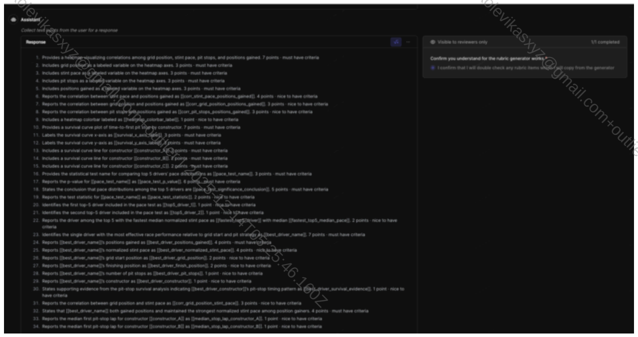
[redacted]



Collect text inputs from the user for a response

Response

Rubric Criterion	Type	Rating	Decision	Justification
Provides a heatmap showing correlations between grid position, stint pace, pit stops, and positions gained.	Must Have	–	IGNORED	This requires visual/plot comparison, excluded.
Heatmap is semantically the same as the reference.	Must Have	–	IGNORED	This requires visual/plot comparison, excluded.
Reports Spearman's ρ between grid position and stint pace (normalized) as −0.02.	Must Have	5	PASS	The model correctly reports the Spearman's ρ as -0.02.
Reports Spearman's ρ between grid position and pit stops as +0.53.	Must Have	5	PASS	The model correctly reports the Spearman's ρ as +0.53.
Reports Spearman's ρ between grid position and positions gained as +0.36.	Must Have	5	PASS	The model correctly reports the Spearman's ρ as +0.36.
Reports Spearman's ρ between stint pace (normalized) and pit stops as +0.74.	Must Have	5	PASS	The model correctly reports the Spearman's ρ as +0.74.
Reports Spearman's ρ between stint pace (normalized) and positions gained as −0.41.	Must Have	5	PASS	The model correctly reports the Spearman's ρ as -0.41.
Reports Spearman's ρ between pit stops and positions gained as −0.20.	Must Have	5	PASS	The model correctly reports the Spearman's ρ as -0.20.
Reports Kruskal–Wallis H statistic as 22.996 with p-value as 1.27×10^{-4} for pace distribution differences among top-5 finishers.	Must Have	5	PASS	The model correctly reports the H statistic as 22.996 and the p-value as 1.27×10^{-4}.
Reports Mann–Whitney p-value for Leclerc vs Russell as approximately 0.0416.	Must Have	5	PASS	The model correctly reports the p-value for Leclerc vs Russell as approximately 0.0416.
Reports Mann–Whitney p-value for Verstappen vs Russell as approximately 0.0466.	Must Have	5	PASS	The model correctly reports the p-value for Verstappen vs Russell as approximately 0.0466.
Provides a survival curve of pit-stop timing by constructor.	Must Have	–	IGNORED	This requires visual/plot comparison, excluded.
Survival curve is semantically the same as the reference.	Must Have	–	IGNORED	This requires visual/plot comparison, excluded.
Reports median first-stop lap for Alpine, Aston Martin, Ferrari, and Kick Sauber as 12–13 laps.	Must Have	5	PASS	The model correctly identifies these constructors and their median first-stop lap as 12–13.
Reports median first-stop lap for Mercedes and Williams as 13 laps.	Must Have	5	PASS	The model correctly identifies these constructors and their median first-stop lap as 13.
Reports median first-stop lap for Haas F1 Team, McLaren, Racing Bulls, and Red Bull Racing as 13.5 laps.	Must Have	5	PASS	The model correctly identifies these constructors and their median first-stop lap as 13.5.



Collect text inputs from the user for a response

Response

1. Provides a heatmap visualizing correlations among grid position, stint pace, pit stops, and positions gained. 7 points · must have criteria
2. Includes grid position as a labeled variable on the heatmap axes. 3 points · must have criteria
3. Includes stint pace as a labeled variable on the heatmap axes. 3 points · must have criteria
4. Includes pit stops as a labeled variable on the heatmap axes. 3 points · must have criteria
5. Includes positions gained as a labeled variable on the heatmap axes. 3 points · must have criteria
6. Reports the correlation between stint pace and positions gained as [[corr_stint_pace_positions_gained]]. 4 points · nice to have criteria
7. Reports the correlation between grid position and positions gained as [[corr_grid_position_positions_gained]]. 3 points · nice to have criteria
8. Reports the correlation between pit stops and positions gained as [[corr_pit_stops_positions_gained]]. 3 points · nice to have criteria
9. Includes a heatmap colorbar labeled as [[heatmap_colorbar_label]]. 1 point · nice to have criteria
10. Provides a survival curve plot of time-to-first pit stop by constructor. 7 points · must have criteria
11. Labels the survival curve x-axis as [[survival_x_axis_label]]. 3 points · must have criteria
12. Labels the survival curve y-axis as [[survival_y_axis_label]]. 3 points · must have criteria
13. Includes a survival curve line for constructor [[constructor_A]]. 2 points · must have criteria
14. Includes a survival curve line for constructor [[constructor_B]]. 2 points · must have criteria
15. Includes a survival curve line for constructor [[constructor_C]]. 2 points · must have criteria
16. Provides the statistical test name for comparing top 5 drivers' pace distributions as [[pace_test_name]]. 3 points · must have criteria
17. Reports the p-value for [[pace_test_name]] as [[pace_test_p_value]]. 6 points · must have criteria
18. States the conclusion that pace distributions among the top 5 drivers are [[pace_test_significance_conclusion]]. 5 points · must have criteria
19. Reports the test statistic for [[pace_test_name]] as [[pace_test_statistic]]. 2 points · nice to have criteria
20. Identifies the first top-5 driver included in the pace test as [[top5_driver_1]]. 1 point · nice to have criteria
21. Identifies the second top-5 driver included in the pace test as [[top5_driver_2]]. 1 point · nice to have criteria
22. Reports the driver among the top 5 with the fastest median normalized stint pace as [[fastest_top5_driver]] with median [[fastest_top5_median_pace]]. 2 points · nice to have criteria
23. Identifies the single driver with the most effective race performance relative to grid start and pit strategy as [[best_driver_name]]. 7 points · must have criteria
24. Reports [[best_driver_name]]'s positions gained as [[best_driver_positions_gained]]. 4 points · must have criteria
25. Reports [[best_driver_name]]'s normalized stint pace as [[best_driver_normalized_stint_pace]]. 4 points · nice to have criteria
26. Reports [[best_driver_name]]'s grid start position as [[best_driver_grid_position]]. 2 points · nice to have criteria
27. Reports [[best_driver_name]]'s finishing position as [[best_driver_finish_position]]. 2 points · nice to have criteria
28. Reports [[best_driver_name]]'s number of pit stops as [[best_driver_pit_stops]]. 1 point · nice to have criteria
29. Reports [[best_driver_name]]'s constructor as [[best_driver_constructor]]. 1 point · nice to have criteria
30. States supporting evidence from the pit-stop survival analysis indicating [[best_driver_constructor]]'s pit-stop timing pattern as [[best_driver_survival_evidence]]. 1 point · nice to have criteria
31. Reports the correlation between grid position and stint pace as [[corr_grid_position_stint_pace]]. 3 points · nice to have criteria
32. States that [[best_driver_name]] both gained positions and maintained the strongest normalized stint pace among position gainers. 4 points · must have criteria
33. Reports the median first pit-stop lap for constructor [[constructor_A]] as [[median_stop_lap_constructor_A]]. 1 point · nice to have criteria
34. Reports the median first pit-stop lap for constructor [[constructor_B]] as [[median_stop_lap_constructor_B]]. 1 point · nice to have criteria

Confirm you understand for the rubric generator works. *

● I confirm that I will double check any rubric items which I will copy from the generator

NOTE: Do not edit tasks or make any selection on the “Task Action” field

General Grading Instructions (How the 1-5 scale is used)

1	**General Grading**
	Grade to the lowest dimension across all rubrics (e.g. if instruction following is a 2, the task should be rated a 2)
	~~Each prompt is considered a separate “turn”~~
	~~If no turns fail, grade to the lowest turn for across all turns (e.g. if turn 2 is a 3, the task should be rated a 3) **(09/28)**~~
	~~If BOTH turns meet any criteria under 1-2 Fail, the task is a fail. **(09/28)**~~
	~~If ONLY ONE turn meets any criteria under 1-2 Fail, the task is a 3. **(09/28)**~~
	If the task does not fail and it meets criteria for a 3-4 [Not-Fail] on any dimension, then the entire task must be a 3-4 [Not-Fail]
	All dimensions must be a 5 for the task to receive a 5.
2	**Choosing 1 vs 2 or 3 vs 4**
	When deciding between a 1 or 2, select a 1 if the attempter put little to no effort
	When deciding between a 3 or 4, use your best judgement on how serious you think the minor issue affects the quality of the task.
3	**Prompt instructions or task instructions should always take precedence over other dimensions**
	For example: if the task instructions asks the user to intentionally make spelling mistakes in the prompt, spelling errors in the prompt would not be marked towards a fail.

Grading Rubrics

Dimension	Sub- Dimension	Notes for Auditors	1 - 2	3 - 4	5

Prompt	**Unique Ground Truth** **UPDATED 10/13**	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	- 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	- The answer to the prompt is not time dependent
~~**Prompt**~~	~~**Reasoning Requirement**~~ **REMOVED 10/21**	~~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 failures 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.~~	~~- The prompt requires reasoning in-line with the assigned domain/expertise level, beyond just simple recall~~
~~**Prompt**~~	~~**Prompt Independence**~~ **REMOVED 10/13**	~~All the prompts should be independent from each other. The prompts DO NOT make up a single conversation, so context should not be borrowed from one prompt to another. If you can't answer any prompt without knowing the other prompts, they are not independent.~~	~~- **[Fail - Dependent Prompts]** The prompts provided can't be answered independent from each other~~	~~N/A~~	~~All the prompts provided are independent from each other~~
~~**Prompt**~~	~~**Prompt Open/Close-Ended**~~ **REMOVED 10/08**	~~Note that both the prompts should be independent from each other. These prompts don't make up a single conversation~~	~~The conversation doesn't have at least 1 open-ended and 1 closed-ended prompt~~	~~N/A~~	~~The conversation has at least 1 open-ended and 1 closed-ended prompt~~
~~**Prompt**~~	~~**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~~	~~All the prompts require a plot/graph/chart~~

	REMOVED 09/29				
Prompt	**Prompt Clarity and Specificity**		**[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	- 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	- The prompt is completely actionable by an LLM or chatbot - The prompt contains no conflicting instructions/statements
Prompt	**Open-Ended or Close-Ended Label** **ADDED 10/08**	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	The prompt is correctly labeled as either open-ended or close-ended.
Prompt	**Complexity Label** **ADDED 10/08**	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	The prompt is correctly labeled as either easy, medium, or hard.

				not correct	
Dataset	**Dataset Integrity** **UPDATED 10/21**	**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.~~ **(10/21)**	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**	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, ignoring instructions, flawed reasoning, or use of the wrong files. Presentation issues alone do not qualify as valid failures. There should be at least one critical criteria in the rubric that the SOTA fails on.	**[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	- 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. See the next columns to understand the thresholds. 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. Note: (Do not fail "negative-criteria" as counter-productive if they are phrased correctly and	Use the number of criteria that the CB wrote as the denominator while calculating % values. See the additional notes section for the numerator. Do NOT double count criteria while tallying even if it has multiple issues. - **[Fail - 10%+ Major Rubric Errors]** More than 10% of the criteria contain major issues - **[Fail - 15%+ Moderate Rubric Errors]** More than 15% of the criteria contain moderate or major issues - **[Fail - 25%+ Minor Rubric Errors]** More than 25% of the criteria contain minor or moderate or major issues	Use the number of criteria that the CB wrote as the denominator while calculating % values. See the additional notes section for the numerator. Do NOT double count criteria while tallying even if it has multiple issues. - **[Non-Fail - Up to 10% Major Errors]** Up to 10% of the criteria contain major issues - **[Non-Fail - 5-15% Moderate Errors]** Between 5 and 15% of criteria contain moderate or major issues - **[Non-Fail - 10-25% Minor Errors]** Between 10 and 25% of distinct criteria contain minor or moderate or major issues - **[Non-Fail - Broad but Ratable Criteria]** More than 25% of	The rubric covers the core instruction The rubric does not contain objective inaccuracies The rubric covers all prompt constraints The rubric is specific and relevant If needed the rubric provides direct answers The rubric provides binary criteria The rubric covers all the necessary criteria to create a perfect response. The rubric criteria weights accurately reflect their importance to a good response At least 1 criterion contains a plot plot/chart/graph There are at least two additional criteria describing how the plot/chart/graph should look There are negative criteria for each model failure

		properly given a negative weight (e.g., “The response X [-10]”) and X is something the model does incorrectly). 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). This can include criteria that explicitly contain a result, or outcome, in direct response to a prompt's ask. Or, this can include criteria that respond to implicit, but objectively necessary, steps to satisfy a direct ask from the prompt. Missing criteria should fall into one of 3 buckets: - Addressing safety of the response - Addressing accuracy of the response - Assessing instruction following of the response (Are all direct requests covered?) **(10/21)** To avoid rubrics that are extremely long, CBs should include at least 20% of the required rubrics if a single request would result in more than 10 rubric items. Count as missing criteria only if this threshold is not met. - Example: “List the top 30 countries that…” - In this case, only a minimum of 6 countries would be needed, however these items should be randomly selected, not all in order such as 1st, 2nd, 3rd, etc - If the CB only wrote 4 criteria, that counts as 2 missing criteria **(10/27)** Prompt requests for items that *can* be broken up into many components and checked individually, but are by nature a single object (e.g., matrices) can be covered by a single rubric item and do not necessarily need to be spot-checked entry-wise. - Example: “Calculate a matrix of the pearson correlations between variables x1, x2, x3…x100” - Response: {100x100 correlation matrix} - A single criterion assessing the matrix as a whole is acceptable (instead of a rubric that checks 20% of the elements, which would total ~2,000 criteria in length) 3. [Major Imbalance] - There is an obvious and egregious misalignment in the magnitude of the assigned weight of the		criteria are broad but most people would be able to determine if they were met (e.g., “the response should be humorous”) - **[Non- Fail - Minor Imbalance]** More than 25% of the rubric criteria weights are roughly correlated with their importance, with some room for debate but no severe errors. - **[Non-Fail - Non-Random Rubric Selection]** CB followed the 20% suggestion for prompt requests that would result in more than 10 criteria (see [Missing Criteria] for context), but their selection of rubrics is not random at all **(10/21)**	

criterion against its actual importance to the prompt. (e.g., the criterion is weighted +15 when it should reasonably be rated +2)
- Positive weights are given to rubric criteria when they should have negative weights, or vice versa.
- Please check the notes below to understand more about weights and do NOT flag minor weight imbalances here.

4. [Not Self-Contained]
- The criterion has a GTFA but does not provide the answer. The criterion has vague language (e.g., “should be accurate”), missing expected outcomes/ examples where relevant. The criterion does not contain all the information needed to evaluate a response independent of the prompt.

E.g.
❌ Mentions the smallest moon of Saturn.
✅ Mentions the smallest moon of Saturn is Aegaeon.

Note: Not Self Contained means it does not contain all the information needed to evaluate a response without relying on basic domain knowledge with negligible reasoning (definition comprehension and trivial equivalence identification)

5. [Plot-based criteria]
- There should be a criterion that checks that the plot produced by the response is "semantically same" as the expected plot. This rubric should also have an image attached to it.
- If this criterion or its corresponding image is missing, count that as 1 major issue

----

MODERATE ISSUES:

1. [Atomicity]
- Rubric contains unrelated prompt requests combined into one rubric criterion

		2. [Overlapping Criteria] - Criteria that evaluate the same exact aspect of the response - Each set of overlapping criteria can only be counted as one failing criteria contributing to a “moderate issue” - How to think about overlap w.r.t negative criteria? **(10/21)** ● There’s always going to be some amount of overlap between the negative criteria and the corresponding positive criteria. Consider the example prompt that asks for the best soccer player. Let’s assume the answer is Messi. ● “[+8] States that the best soccer player is Messi” and “[-8] States that the best soccer player is Ronaldo” ← this is okay ● “[+8] States that the best soccer player is Messi” and “[-8] Does not state that the best soccer player is Messi” ← this is not okay as the negative criteria is just the inverse of the positive ● In essence, penalizing common mistakes is good. But if the negative criteria is just the direct negation of the positive criteria, that’s not adding anything of value and is treated as an overlapping criteria ● Two negative criteria can overlap among themselves too, and we should evaluate it similar to how we evaluate overlap between two positive criteria - This includes positive/negative criteria that are redundant (covering the same aspect) -**09/27** - This set of redundant criteria can only be counted as one failing criteria contributing to a “moderate issue”: [+15] Identifies New Delhi as the capital of India | [-10] Identifies Mumbai as the capital of India -**09/27** 3. [Vague Criteria] The criterion is too vague or abstract to be meaningfully evaluated, making it impossible to determine whether it has been met. 4. [Plot-based criteria] **(10/21)** - Along with a criterion that checks if the plot generated is semantically the same, there should be at least one additional additional criterion describing how the plot/chart/graph should look. For example "The line chart shows an increasing trend"			

		- If there are zero such additional plot criteria, count that as one "moderate issue". If there is at least one, count as no issues. --- **MINOR ISSUES:** 1. [Binary] Criterion is not binary (true or false) or objective (a majority of readers should agree on whether a given model response satisfies the criteria). Even if the criterion must rely on some level of personal interpretation, it should be something that >75% of people would rate the same way. 2. [Double Negative] - A negative criterion flags something absent from the response, instead of something present E.g. ❌[-5] The response fails to correctly calculate the proportion of debt funding as 92% ✅[+5] The response correctly calculates the proportion of debt funding as 92% 3. [Unnecessary Criteria] The criterion checks for something that adds no value to the response. If there is any reasonable argument for the inclusion of the criterion, the criterion should NOT be counted as an issue for this condition. Give the CB the benefit of the doubt because this is inherently subjective. Note: Criteria that relate to the stylistic/formatting elements of the response should not be considered unnecessary if they support the CB's preference for visually pleasing responses 4. [Negative phrasing] The criterion is phrased negatively with a positive weight, when it could've been phrased positively with a negative weight. ([+10] Does not list pink as a rainbow color **vs.** [-10] Lists pink as a rainbow color) -**09/27**			
Rubric Criteria	**Presence of Negative**	Negative criteria are criteria that have negative weights. They should still be written with a positive framing.	N/A	**[Non-Fail - No Negative Criteria]** There are no negative criteria in the rubric.	There is at least one negative criteria in the rubric

	Weighted Criteria	For example, "States that 2+2 = 5" with a weight of -20.			
~~**Rubric Criteria**~~	~~**Presence of Non-Outcome Criteria**~~ **REMOVED 09/29**	~~The rubric should *not* include any criteria that do not target the final generated response (e.g., Reasoning/Process criteria).~~	~~**[Fail - Has Non-Outcome Criteria]**~~ ~~- The rubric includes one or more criteria that do not target the generated response (Outcome).~~	~~N/A~~	~~- All of the rubric's criteria target the generated response (Outcome).~~
Quality Control Notes	**Clarity**	Quality Control Notes: Comprehensive explanation of how the ideal workflow should look in order to obtain the prompt's answer.	N/A	- The Quality Control Notes have issues, they are either misaligned with the prompt's intent, incorrect or lacking details	- The Quality Control Notes are aligned with the prompt's intent
All CB Generated Content	**Unlisted Minor Errors**		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.	- There are no unlisted errors that you feel degrade the quality of the task

Appendix

Understanding Weights:

- Each rubric criterion must have a weight between -40 and 40, based on its importance to the prompt.
- Accuracy-related criteria (e.g., factual correctness, calculations, results) should receive the highest weights, since they are the most critical.
- Instruction-following criteria (e.g., formatting, phrasing, or minor task requirements) should be assigned lower weights than accuracy but still enough to matter
- During training, all the weights are normalized, so the absolute weight values are not important as much as the relative weight values across the rubric. It's only important that the weights are proportionate to the importance of the rubric. If a CB chooses

to rate all criteria between -5 and 5, that's okay as long as they're proportional.

Understanding Negative Weights:

Rubrics need to give a comprehensive toolkit for grading the ideal model response. For the cases when the experts know about “common error patterns” they can create rubric items with negative weights to penalize the common errors.

For the purposes of an example let’s think that our prompt is “Give me the 7 standard colors of rainbow. What are they?”
The rubric for this would be something like:

None

```
[+5] states that the first color is red
[+5] states that the second color is orange
[+5] states that the third color is yellow
.
.
[-10] states that pink is a rainbow color
```

In this example, we know that pink can be confused as a rainbow color and are writing a negative criteria to penalize it.

To understand scoring - think about a teacher grading an essay, where every time they locate the statement in the essay they grant the corresponding points.So, the grading logic is:

- Is this statement true when you read it against the response?
- If True → add the points without changing the sign (which would mean that negatively weighted criteria are subtracted from score)
- If False → do nothing / 0 points

For the above example, if the model response was “The first rainbow color is red, the second rainbow color is orange and the third rainbow color is pink …”
We would grade:

- +5 because the first criterion is present (True)
- +5 because the second criterion is present (True)
- + (-10) because the criterion for pink is also present (True)