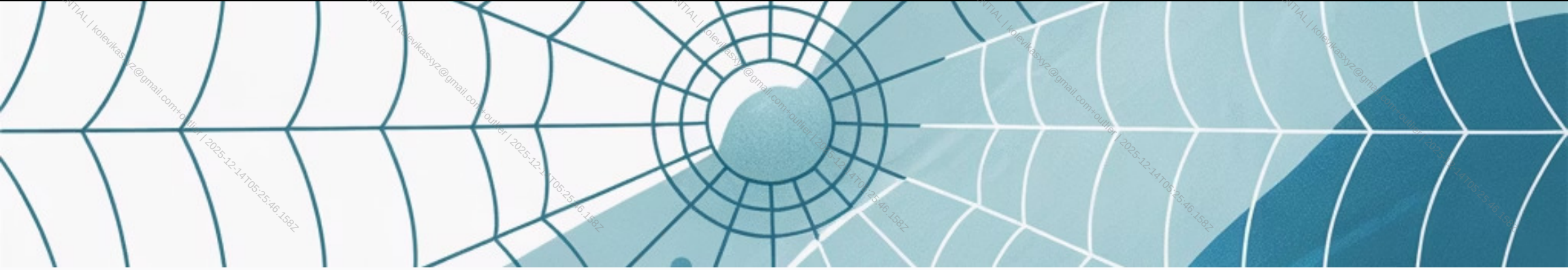


Spider Web Rubrics

Welcome!

Project objectives

In this project, your task is to design prompts that challenge AI models to analyze, manipulate, reason with complex data, and generating a plot or visualization,

The prompts should be complex enough to stump the model.

For each prompt, you will create clear rubric statements that will be used to check both the numerical outputs of the answers to the prompts and the correctness of the plot.

Understanding the Tasking Workflow Structure

The tasking workflow follows a structured three-level progression designed to ensure quality at every stage of prompt and rubric development.

01

Attempt Level: Prompt Creation & Quality Control Notes

When you first join the project, you'll receive Attempt permissions. Your primary responsibility is creating complex prompts focused on Data Analysis, then adding detailed Quality Control Notes. The QMs and Project Team will review your work to ensure it meets all high-quality prompt requirements.

02

L0: Rubric Creation

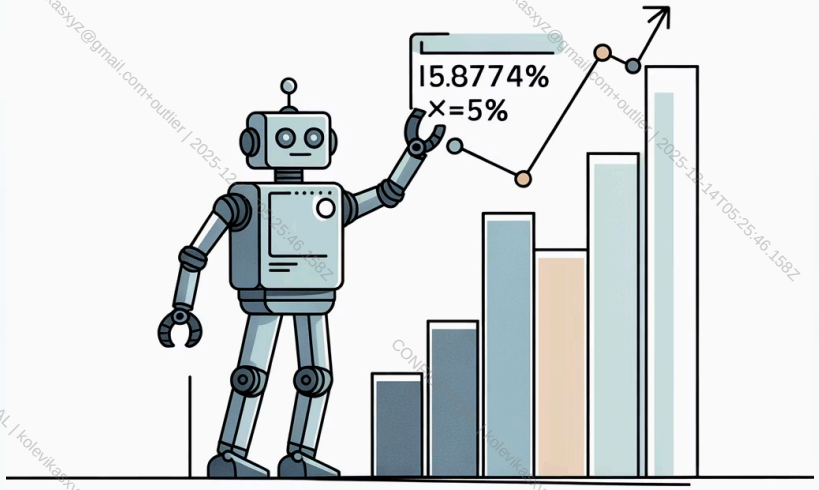
After a prompt is approved, you advance to creating the comprehensive Rubric for evaluation. This involves developing specific criteria that judges will use to assess model responses against your prompt requirements.

03

L1: Review Level

If given Reviewer status, you'll be responsible for reviewing tasks end-to-end—from initial prompt creation through Quality Control Notes and final Rubric validation, ensuring consistency and quality across all project submissions.

What Constitutes a Model Failure?



Spider Web focuses on Data Science tasks, useful failures involve **numerical values or analytical reasoning that results in wrong outputs in terms of values and explicit claims**.

Failing the model for using the wrong plot color or forgetting to round a number is not meaningful—these are instruction-following issues, not core Data Science failures.

Valid model failures are: **incorrect calculations, flawed statistical analysis, or misinterpreting data relationships** creates valuable training data to make the model stronger.

Important: Model crashing is not considered a valid model failure for this project.

Understanding Model Responses

When we say **model response**, we don't only mean the answer currently shown in your task interface. Instead, we mean the response the model will generate to your prompt **during training**—which could happen tomorrow, six weeks from now, or six months later. The model generates the response it predicts is correct at that point in its training.

Why This Matters

Timelessness

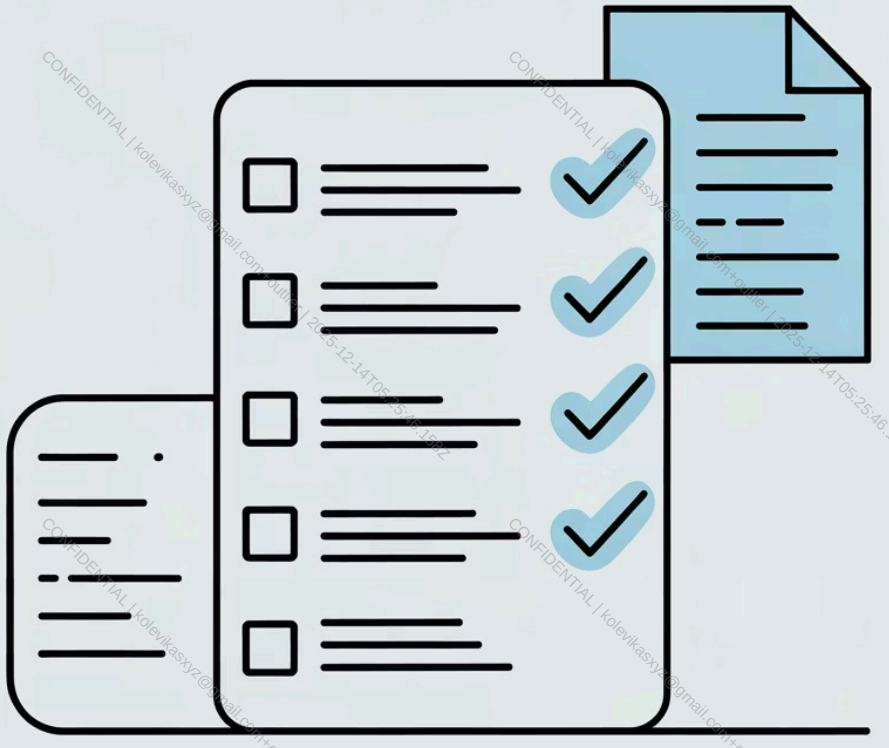
Prompts must be written so their answers don't change over time. If the answer is time-sensitive (or the code changes outputs every time you run it) and the customer trains on your data months later, the judge will never pass the model's response—rendering the training data useless.

Single Ground Truth Final Answer

Prompts should be designed with only one correct final answer in mind. If multiple answers are possible, the model may generate a different (but valid) answer than the one your rubric addresses, again reducing data utility. NOTE: for open ended prompts, we do allow some flexibility as two experts may give two slightly different answers.

Visibility for Judges

In this project, Judges see the final output. If a rubric asks judges to evaluate something they can't see (like the model's thinking process), the model will never pass evaluation.



Critical Task Requirements

Every task comes with specific requirements you must follow carefully to ensure your work meets project standards.

Mandatory Requirements

- **Complexity** – Prompts in this project must be of a HARD complexity level.

Encouraged Guidelines

- **Suggested Dataset** – This is the recommended data for your task. You need to double check that these meet the minimum requirements.

Dataset Requirements & Resources

Accepted file formats include CSV, JSON, XLSX, JSONL, TSV, and XML. **You must attach the files in the first prompt**.

You will be provided with datasets in your tasks.
If the datasets do not meet the project requirements, you can find other datasets in the following recommended resources:

- **Google Dataset Search** - Comprehensive search engine for datasets
- **Kaggle Datasets** - Extensive collection of community-contributed data
- **Tableau Public Data Sets** - Curated public datasets
- **Data.gov Catalog** - U.S. government open data
- **DataHub Collections** - Organized dataset collections
- **NASA Earthdata** - Scientific research datasets

You'll need to download each dataset and save the source links from the websites you used for documentation purposes.

Dataset Submission Requirements

Always double check for the following:

1

Minimum Dataset Count

You must use **at least 3 datasets per task**. All three must be attached when submitting your prompt, and you must paste the source links to all datasets used.

2

Accessibility & Licensing

Datasets must be **publicly accessible** (no paywalls) and **open source** (free for use, no private licenses required).

3

Language & Structure

All datasets must be **in English**, **clean and structured** with headers on the first row, and free from corruption or formatting issues.

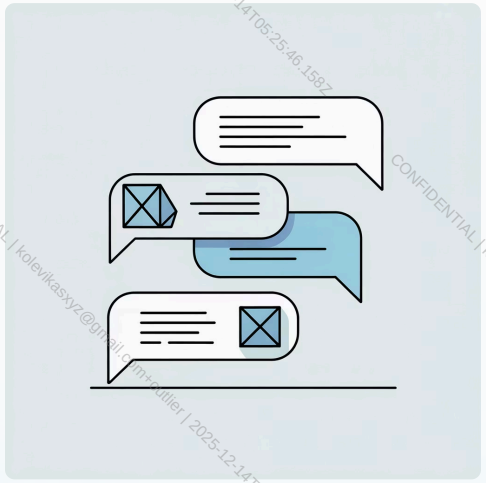
4

Size Guidelines

Datasets should be **sufficiently large**—aim for at least 5 columns and 100 rows (not a hard requirement). Total size across all 3 datasets must be **less than 1 GB**.

Reminder: Double-check each dataset before submitting to ensure it opens correctly and meets all requirements. Invalid or inaccessible datasets will result in task rejection.

Using ChatGPT-5 for Model Responses



🚨 **CRITICAL REQUIREMENT**

For this project, you are required to use ChatGPT-5 PLUS (chatgpt.com) to generate all model responses. This means you must upload the necessary files directly into ChatGPT-5 and complete the entire workflow inside that platform—not in other models like GPT-4o or GPT-3.5.

Access & Reimbursement

Plus Plan Required

You'll need to upgrade to the **Plus Plan** to unlock ChatGPT-5 access. Visit ChatGPT Pricing for subscription details.

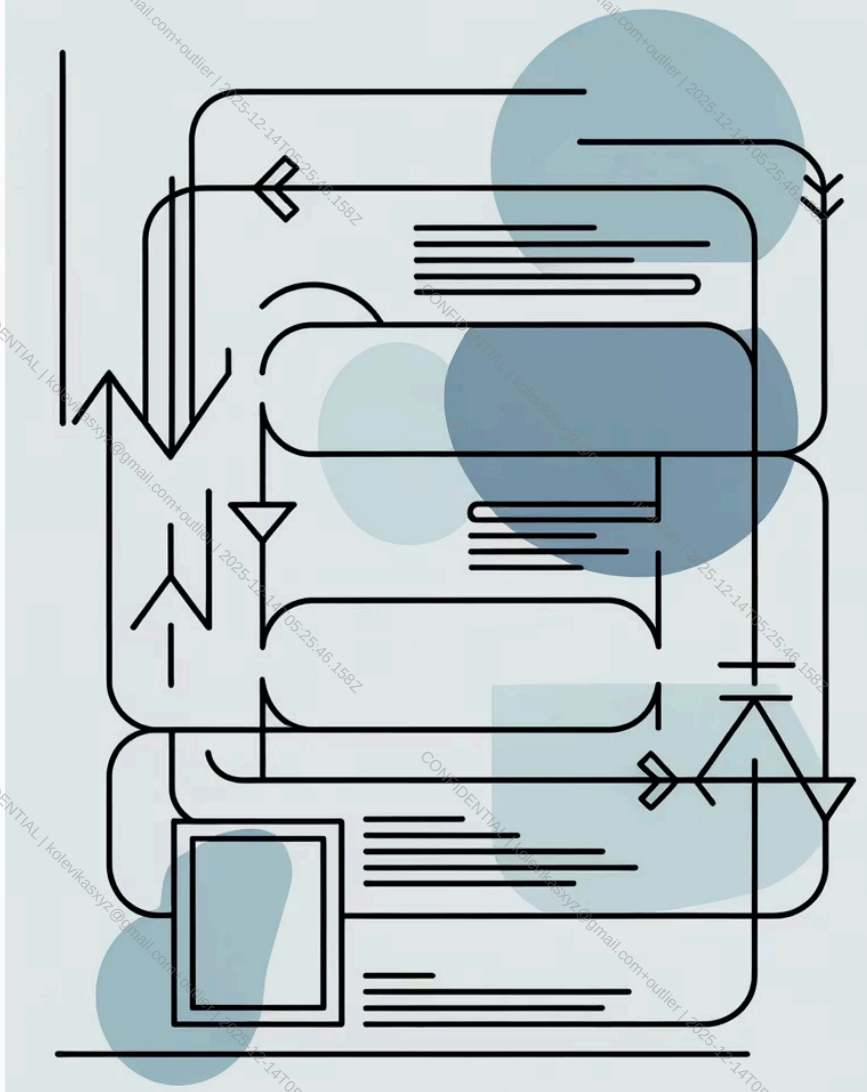
Reimbursement Conditions

The cost of the Plus Plan will be **reimbursed**, but only if you submit the correct ChatGPT-5 conversation link and use the correct model. Using any other model (4o, 3.5, etc.) will disqualify you from reimbursement.

One-Time Benefit

This reimbursement will only happen **once per cicle**. Double-check that you've selected the correct model before starting your work.

Tasking Workflow



Creating Effective Prompts

You'll write your prompt in the ChatGPT-5 platform. The prompt must follow the assigned complexity level and adhere to comprehensive Prompt Guidelines.

Critical Considerations

Create a New Chat for Every Prompt

Always start fresh conversations to ensure clean context and avoid contamination from previous interactions.

Upload At Least 3 Datasets

Attach your three datasets (or more) before asking prompts focused on Data Analysis or Data Science topics.

Use All the Uploaded Datasets

Your prompts must incorporate and reference all datasets uploaded, not just a subset.

Prompt-Dataset Dependency

Every prompt must rely on the attached datasets. Questions should not be answerable without using the dataset information.

Prompts Must Stump the Model

Design challenging prompts that expose model weaknesses in data analysis and reasoning.

Prompt Writing Guidelines

Specific & Well-Structured

Prompts should be clear, detailed, and organized. They may include multiple related questions requiring the model to read, manipulate, and reason through data, as well as run commands to generate results.

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.

Hard Complexity Required

Prompts must be COMPLEX—meeting Hard Complexity standards. See complexity guidelines in project documentation for specific requirements.

Include Plot Generation

Include at least one question that requires the model to generate a plot, chart, or visualization from the data.

Ensure Timelessness

Prompts should not depend on execution time, so ensure the ideal response does not change with time or current events.

Open & Close-Ended Questions

Create prompts based on task assignment: closed-ended (all experts arrive at same answer) or open-ended (experts could provide different but valid answers).

Important: Only ONE prompt needs to be submitted to GPT-5. Follow-up prompts are **not allowed** in this project.

Identifying Model Failures

To proceed with the task, **the response must contain a significant failure**—such as an incorrect answer, use of wrong files, flawed reasoning, or miscalculation. Presentation issues or model crashing do not qualify as valid failures.

When the Model Doesn't Fail

If the model doesn't fail, you'll need to go back and create another prompt that's more complex than your previous attempt. Do not write follow-up prompts to guide the model—start fresh.

Verify Your Prompt Clarity

If you feel the model is giving a completely nonsensical answer, check that your prompt is clear and not ambiguous. Fix it and re-prompt the model in a new conversation.

Pro-Tips for Finding Failures 💡

- **Check for factual inaccuracies** in calculations or data interpretations
- Look for **hallucinations in the data** (made-up values or statistics)
- Identify **missing steps or misleading steps** in the solution process **that results in wrong values**.
- Verify the model response **completely fulfilled your prompt** requirements
- Confirm **correct file usage** and appropriate dataset references

You'll be asked to document all model failures you found with brief explanations justifying each error.

Writing Quality Control Notes

You'll need to provide a comprehensive explanation of how the ideal workflow should look to obtain the prompt's answer. This explanation is essential for reviewers to validate correctness and understand your analytical approach.

Required Components

1

Prompt Goal

Clearly state what the prompt is asking for. No need to copy/paste the full prompt—just summarize the objective concisely.

2

Step-by-Step Solution

A detailed breakdown of how to solve the prompt, covering every question and requirement. Reference specific formulas, column names, and sub-steps.

3

Code Blocks

Include code for plots, calculations, or other operations. All code must be properly formatted in code blocks.

Quality Control Notes Template

Prompt Goal:
[Summarize the main objective of the prompt in your own words.]

Step-by-Step Solution:
[List out all steps required to solve the prompt, including formulas, operations, reasoning, and code blocks to obtain the plots/graphs. Ensure every requirement from the prompt is covered.]

Code Used to Obtain Answers:
[Paste all relevant code here, formatted in triple backticks. This should include code for analysis and visualization.]

Documenting Final Outputs

You'll be asked to provide the final outputs from your prompt in a dedicated field. This is where you write the ground-truth final answers.

What to Include

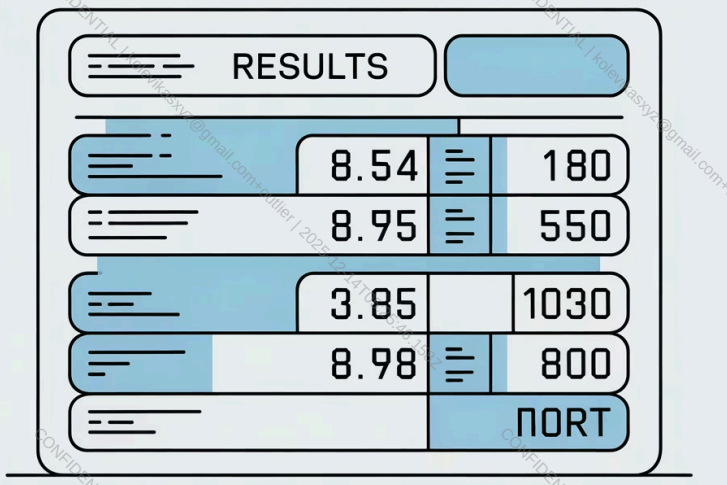
Please only include the final outputs of the questions in your prompt—that is, the ground-truth final answers without explanations, code, or additional context.

Example Format

If the prompt asks for the Pearson correlation, the average population, and the GDP year-over-year, what you should enter is:

- Pearson correlation = 0.04
- Average population = 2,455,000
- GDP YoY = 0.343

> **Avoid adding explanations, code, or any other content beyond the required output.** If the prompt does not explicitly ask for it, do not include it.



Rubrics

A rubric is a comprehensive checklist of criteria that defines an ideal response to a user prompt. You can think of a rubric as a complete, non-redundant, and accurate set of truth conditions specifying under which conditions a response to a given prompt is a good-quality response.

Understanding Rubric Criteria

Each criterion is an item on the checklist that defines a single condition for the ideal response. The set of all criteria should consist of individually necessary and jointly sufficient items for the response to be perfect.

Rubric Quality Standards

Criteria MUST BE:

- **Appropriately atomic** – Related to a single discrete challenge in the prompt
- **Specific** – Containing sufficient detail, including answers/examples that don't leave it completely open-ended
- **Accurate** – Correct by fact check and logical reasoning
- **Categorized** – All items should fall under appropriate categories

A Rubric MUST BE:

- **Comprehensive** – No missing steps for an ideal response
- **Self-contained** – Sufficient details/examples provided for specificity
- **Relevant** – No unnecessary or unrelated criteria
- **Non-Repeating** – No redundant steps; don't penalize the same mistake twice
- **Reflective** – Reflects all explicit asks of the prompt

Rubric Writing Principles in Spider Web

When writing rubrics, follow these principles to ensure clarity, precision, and coverage of all key requirements.

Plot-Based Criteria

Every rubric MUST have at least 1 criterion that requires a plot: Include the correct reference plot and both verbal criteria (descriptions) and visual reference to check semantic equivalence: "**Displays a bar plot that is semantically the same as the attached plot.**" Separate criteria should differentiate the plot view from content requirements: "**Includes a title "Projected vs Actual values" in the bar plot**.

Precision & Specificity

Always be highly specific. ✅ Example: "Number must be 2.342" ❌ Avoid: "Number must be 2". Precision prevents ambiguity and ensures consistent evaluation.

Content Coverage: Criteria Should Focus on OUTPUTS, not processes

Rubrics must mention: factual answers to questions, explicit instructions from the prompt regarding outputs. For every output failure identified in the model response, create a criterion addressing it.

Style & Wording

Rubric criteria must start with simple present tense (e.g., "States that...", "Includes a...", "Provides a...", "Mentions that..."). ✅ "Calculates a p-value of p=0.0001" ❌ "The response must calculate the p-value"

Categorization

Classify items as: **Must-Have (Essential)** – Direct instruction following, explicit asks from the prompt. **Good-to-Have (Nice to Have)** – Reasoning quality, factual correctness beyond explicit asks, implicit or inferred instructions.

.

Understanding Negative Criteria

What are Negative Criteria?

A negative rubric item is a criterion that deducts points for a common mistake, omission, or incorrect reasoning. It identifies what should not be done and assigns a negative weight, ensuring clear penalties for errors.

Why are they Important?

They highlight common mistakes to avoid, ensuring wrong or missing information is fairly penalized. This creates balanced grading, rewarding correct answers while penalizing errors, leading to accurate and consistent results. They make feedback transparent, helping models learn from and avoid repeating mistakes.

Applying Negative Criteria Effectively

Understanding the role and importance of negative criteria is the first step. To ensure these criteria are applied fairly and consistently, a clear set of rules and guidelines is essential. These principles help maximize the benefits of negative rubrics while maintaining accuracy and transparency in evaluation.

Prevent Double Penalization

Make sure you are not creating a rubric item that is closely covered under a positive rubric item. If we made the negatives simple opposites, we'd double-penalize the same issue.

Avoid Polar Opposites

Ensure negative criteria evaluate distinct flaws, rather than merely stating the opposite of a positive criterion. For example: **Positive:** States the best soccer player is Messi. **Negative:** States the best soccer player is not Messi (This is generally to be avoided. Focus on what's incorrectly present, not what's absent.)

Target Common Errors

Focus on typical mistakes or failures that frequently occur in responses. This ensures the rubric is practical and addresses real-world issues rather than hypothetical ones.

Look for "Extra" Claims

Negated criteria can target incorrect claims the model adds that weren't explicitly requested, similar to statements related to the required outputs but included as extra content. Essentially, penalize "extra" hallucinations that we would otherwise consider as "nice-to-have" content if they were correct.

1

Ensure Proportional Weight

Assign weights to negative criteria that reflect the severity and impact of the mistake. Minor flaws should result in smaller deductions, while significant errors warrant larger penalties, proportionate to the overall value.

2

Use Positive Framing

Although evaluating negative aspects, phrase the criteria in terms of an undesirable element *being present*.

3

Minimum Criteria Count

Aim for a minimum of 10 distinct negative items. This ensures comprehensive coverage of potential shortcomings, preventing oversights in assessment.

Rules for Negative Rubrics

The general principles of rubric writing also apply to negative criteria, ensuring fairness and clarity in evaluation. Remember these key guidelines:

Be Specific and Atomic

Each negative item must penalize one clear, observable mistake, not multiple issues at once. Avoid compound statements like "fails to define or explain variable"; split them into distinct criteria if necessary.

Self-Contained Statements

Each criterion should be fully self-contained, providing all necessary information for a judge (human or LLM) to evaluate the model's response without needing external context or prior knowledge of the topic.

Avoid Redundancy

Ensure that negative rubrics do not overlap or test nearly the same aspect. If two items cover similar ground, remove one to maintain efficiency and avoid penalizing the same mistake twice.

Objective Language

Describe what is wrong in factual, checkable terms rather than vague judgments. The language used should be objective and observable, allowing for consistent and unbiased evaluation.

To further refine the application of negative criteria, consider these detailed rules:

Thank You!

Any questions?