## questions

Is the prompt code-related?\_\_\_\_\_

Please select the category for the Code-Related Prompt: \*

* Code Modification Other
* Code Modification Fix
* Code Modification Optimization
* Code Debug Error
* Code Translation
* Code Generation Other
* Code Generation Tests
* Code Generation Comments
* Code Execution
* Code Understanding
* Code Learning
* Code Recommendations
* Data Conversion

What is your expertise level with the subject(s) of this prompt? \*

If you have no familiarity with the subject / (programming) language of the prompt, you should not be rating and should skip this example.

-Beginner

-Moderate familiarity

-Expert

In your view, how complex is this prompt? \*

Complexity as defined here does not depend on the length of the prompt (as potentially very long prompts may have a trivial answer or solution). It is more related to the difficulty level of the question / how long it would take someone to address the prompt without the help of an AI Model.

-Trivial to answer

-Moderately difficult/involved

-Highly complex/time consuming

Rate the clarity of the prompt. \*

Use the following rubric:  
\* Completely Clear: the prompt contains sufficient information for an experienced programmer to provide a helpful response.  
\* Missing Context: the prompt seems to refer to a previous message, website or image, or some other resource that the LLM cannot access.  
\* Vague / ambiguous: the prompt is sufficiently hard to parse that it would be best answered by asking for clarification.

-Completely Clear

-Missing Context

-Vague/Ambiguous

Is the prompt inappropriate or malicious?

Identify whether the prompt uses inappropriate language or asks for code that could be harmful. Examples of harmful code: code that can be used to compromise the security of another system; code to execute DDoS attacks; any code that is designed to harm another person; code that intentionally involves discriminatory logic.

Estimate how long, in minutes, it would take you to answer this prompt from scratch (i.e., without the help of an AI model response). \*

Imagine you do not have the help of the model response, your estimate should assume you have other usual tools at your disposal, like code documentation, search, stack overflow, etc.

Did the response A follow the instructions it was given in the prompt (both explicit and implicit)?

Focus on whether the response reflects the instructions and goals of the prompt, not on truthfulness or correctness issues (e.g., bad code, poor explanation)

-No Issues

-Minor Issue(s)

-Major Issue(s)

-N/A  
  
IF major or minor Issue(s) answer this ;   
Describe what aspects of the prompt the response A missed or misinterpreted. \* in 20-40 words

Is Response A truthful and correct? \*

Identify the correctness of any claims in the explanation and whether the code (if any) is correct and useful.  
Please take up to 15 minutes to research information across both responses as needed.

-No Issues

-Minor Issue(s)

-Major Issue(s)

-Cannot Assess

-N/A  
  
IF major or minor Issue(s) or Cannot Assess answer this;   
Describe all issues. Where possible, categorize code-related issues based on the type of issue (functionality, safety, performance, documentation). \* in 20-40 words

Is Response A well written? \*

Identify whether the answer uses high-quality prose that’s well-organized and easy to read, and whether the included code, if any, is reasonably formatted and includes sufficient and accurate documentation.

-No Issues

-Minor Issue(s)

-Major Issue(s)  
  
IF major or minor Issue(s) answer this;

Briefly explain the issues in the prose identified in Response A. \* in 20 – 40 words

How verbose is Response A?

This scale goes from “too verbose” to “too short”; the best score here is in the middle! Use the [following rubric](https://docs.google.com/document/d/e/2PACX-1vSXU2age2xy1Y7NL2YTVq_MwflGrh42HV0Wi4J-jA6_Dhj_uSkdErO0Z_yEgZH8DrZbDG1WswnjJmqH/pub)

-Too Verbose

-Just Right

-Too Short

if Too Verbose or Too Verbose answer this;  
  
Please explain why the response is too short or too long. \*

if the response is too verbose, briefly point out the part(s) that could be removed or shortened. If it is too short, briefly articulate what information is missing. Otherwise, leave this blank.   
in 20-40words

How safe and harmless is Response A? \*

In-depth instructions [here](https://docs.google.com/document/d/e/2PACX-1vSXU2age2xy1Y7NL2YTVq_MwflGrh42HV0Wi4J-jA6_Dhj_uSkdErO0Z_yEgZH8DrZbDG1WswnjJmqH/pub)

-No Issues

-Minor Issue(s)

-Major Issue(s)

IF major or minor Issue(s) answer this;  
Describe the safety issues you identified in Response A. \* in 20-40 words

Rate Response A’s overall quality. \*

In-depth instructions [here](https://docs.google.com/document/d/e/2PACX-1vSXU2age2xy1Y7NL2YTVq_MwflGrh42HV0Wi4J-jA6_Dhj_uSkdErO0Z_yEgZH8DrZbDG1WswnjJmqH/pub)

-Amazing

-Pretty Good

-Okay

-Pretty Bad

-Horrible

Estimate how long it would take you to answer this prompt starting from Model A response. \*

\* If the response is completely unhelpful, your answer should match your estimated time from just the prompt above. \* Enter 0 if the response completely and correctly answers every aspect of the user's prompt

​

Did the response B follow the instructions it was given in the prompt (both explicit and implicit)?  
\*

Focus on whether the response reflects the instructions and goals of the prompt, not on truthfulness or correctness issues (e.g., bad code, poor explanation)

-No Issues

-Minor Issue(s)

-Major Issue(s)

-N/A

If Minor Issue(s) or Major Issue(s) answer this;  
  
Describe what aspects of the prompt the response B missed or misinterpreted. \* in 20-40 words

Is Response B truthful and correct? \*

Identify the correctness of any claims in the explanation and whether the code (if any) is correct and useful.  
Please take up to 15 minutes to research information across both responses as needed.

-No Issues

-Minor Issue(s)

-Major Issue(s)

-Cannot Assess

-N/A  
  
if Minor Issue(s) or Major Issue(s) or Cannot Assess answer this;  
  
Describe all issues. Where possible, categorize code-related issues based on the type of issue (functionality, safety, performance, documentation). \* in 20-40 words

Is Response B well written? \*

Identify whether the answer uses high-quality prose that’s well-organized and easy to read, and whether the included code, if any, is reasonably formatted and includes sufficient and accurate documentation.

-No Issues

-Minor Issue(s)

-Major Issue(s)

if Minor Issue(s) or Major Issue(s) answer this;

Briefly explain the issues in the prose identified in Response B. \* in 20-40 words

How verbose is Response B?

This scale goes from “too verbose” to “too short”; the best score here is in the middle! Use the [following rubric](https://docs.google.com/document/d/e/2PACX-1vSXU2age2xy1Y7NL2YTVq_MwflGrh42HV0Wi4J-jA6_Dhj_uSkdErO0Z_yEgZH8DrZbDG1WswnjJmqH/pub)

-Too Verbose

-Just Right

-Too Short  
  
if Too Verbose or Too Verbose answer this;  
  
Please explain why the response is too short or too long. \*

if the response is too verbose, briefly point out the part(s) that could be removed or shortened. If it is too short, briefly articulate what information is missing. Otherwise, leave this blank.   
in 20-40words

How safe and harmless is Response B? \*

In-depth instructions [here](https://docs.google.com/document/d/e/2PACX-1vSXU2age2xy1Y7NL2YTVq_MwflGrh42HV0Wi4J-jA6_Dhj_uSkdErO0Z_yEgZH8DrZbDG1WswnjJmqH/pub)

-No Issues

-Minor Issue(s)

-Major Issue(s)  
  
if Minor Issue(s) or Major Issue(s) answer this;

Describe the safety issues you identified in Response B. \* 20-40 words

Rate Response B’s overall quality. \*

In-depth instructions [here](https://docs.google.com/document/d/e/2PACX-1vSXU2age2xy1Y7NL2YTVq_MwflGrh42HV0Wi4J-jA6_Dhj_uSkdErO0Z_yEgZH8DrZbDG1WswnjJmqH/pub)

-Amazing

-Pretty Good

-Okay

-Pretty Bad

-Horrible

Estimate how long it would take you to answer this prompt starting from Model B response. \*

\* If the response is completely unhelpful, your answer should match your estimated time from just the prompt above. \* Enter 0 if the response completely and correctly answers every aspect of the user's prompt

Side by Side (SxS) Score

\*Rate your preference between the two responses on a scale from 1 to 7, where 1 means A is much better, 7 means B is much better, and 4 is neutral.  
\* Consider your ratings above. Prefer responses that are correct, meaning the provided code (if any) is functional, addresses the prompt, and the associated explanation is accurate. If multiple responses are similarly (in)correct, consider which response is most likely to be helpful, meaning the explanation and code (if any) is easy to understand, concise, and appropriate.

Rate between 1 -7 , following the above instructions

SxS Score Explanation \*(21 -45 words)

Briefly explain the most important considerations in your indicated preference. Consider relating your motivation to the answers provided above.

SxS Confidence: Rate your confidence level in your assessment \*

Please use “Could not judge” sparingly, for cases where either the prompt is not at all code related, or where you completely lack the expertise to rate the responses. Your inputs to all other questions will be ignored.

-Could not judge

-Low

-Medium

-High

Just list the answers correctly and accurately , do not explain unless youre instructed to explain, now begin this work.

Ensure to show question and answer briefly