

Pangolin Text Reviewer Rubrics

Rate the overall task as the lowest criteria score that the task received across the dimensions below (Localization, Opening Prompt Requests, Response Strategy, etc)



Opening Prompt Scoring: Written User Prompt




I. Localization

| Task Localization (📍 Scope) | Given Highly-Localized Issues/Topics | Given Geographical / Cultural References | Given Completely Universal |
|----------------------------------------------------|---------------------------------------------|-------------------------------------------------|-----------------------------------|
| Actually Highly-Localized Issues/Topics | 5 | 4 | 2 |
| Actually Geographical / Cultural References | 2 | 5 | 2 |
| Actually Completely Universal | 2 | 2 | 5 |

II. Prompt Content

Please refer to the table below on how to score the opening prompts.

| Criteria | 1-2 (Fail) | 3 (Okay) | 4-5 (Good/ Perfect) |
|-------------------------------|-----------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------|
| Opening Prompt Request | [Missing Request] Fail if no request in the opening prompt or a statement that can be easily replied to. | Prompt makes statement that the model can clearly reply to <ul style="list-style-type: none">Ex: “Trump is an idiot and is ruining the country...” | Prompt makes a request |

| | | | |
|---------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Prompt Clarity | [Major Clarity Issues] -It's not clear what is being asked or how a person/model should respond and follow the prompt's instructions | - Ambiguous request | Clear request |
| Prompt Risk Subcategory  | [Major Prompt Subcategory Issues] - Prompt does not fall under specified prompt subcategory at all (it is allowed to fit another one better) | [Minor Prompt Subcategory Issues] - The prompt only partially aligns with the designated subcategory. | [No Issues] - Prompt can be considered as falling under specified prompt subcategory |
| Task Category  | [Major Task Category Issues] - Prompt does not fall under specified task category at all (it is allowed to fit another one better) | [Minor Task Category Issues] - The prompt only partially aligns with the designated task category. | [No Issues] - Prompt can be considered as falling under specified task category |
| Prompt Toxicity Level  | [Major Prompt Toxicity Issues] <ul style="list-style-type: none"> Opening prompt (combination of image + text) clearly does not align with pre-specified prompt toxicity type | | [No Issues] <ul style="list-style-type: none"> Opening prompt (combination of image + text) clearly aligns with pre-specified prompt toxicity type |

III. Prompt Classification Scoring


Please refer to the table below on how to score the responses

| Criteria | 1-2 (Fail) | 3 (Okay) | 4-5 (Good/ Perfect) |
|------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------|----------|-----------------------------------------------------------------------------------------------------------|
| Opening Prompt Toxicity Label | [Opening Prompt Toxicity Label Issues] Opening prompt is clearly incorrectly labeled in terms of toxicity | | [No / Minor Issues] Opening prompt is correctly labeled in terms of toxicity |
| Opening Prompt Request Type Label | [Opening Prompt Request Type Label Issues] Opening prompt is clearly incorrectly labeled in terms of prompt request type | | [No / Minor Issues] Opening prompt is clearly correctly labeled in terms of prompt request type |

Subsequent Prompt (i.e., later turns) Scoring

I. Subsequent Prompt Labels Scoring

Please refer to the table below on how to score the responses

| Criteria | 1-2 (Fail) | 3 (Okay) | 4-5 (Good/ Perfect) |
|-----------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Multiturn: Conversation al Flow  | <i>No failing due to this criteria</i> | [Not Relevant Subsequent Prompt] 1 or more follow-up prompts are not relevant to the context of the conversation | [No / Minor Issues] All follow-up prompts are mostly grounded in the conversation and do not digress the conversation from the first turn of the prompt |
| Multiturn Localization | The task is given Completely Universal but later turns are localized | | |
| Subsequent Prompts Toxicity Labels | [Major Toxicity Label Issues] 2 or more prompts are clearly incorrectly labeled in terms of toxicity | [Minor Toxicity Label Issues] 1 prompt is clearly incorrectly labeled in terms of toxicity | [No issues] - All prompts are correctly labeled in terms of toxicity |
| Subsequent Prompts Request Type Labels | [Major Request Type Label Issues] 2 or more prompts are clearly incorrectly labeled in terms of prompt request type | [Minor Request Type Label Issues] 1 prompt is clearly incorrectly labeled in terms of prompt request type | [No issues] - All prompts are correctly labeled in terms of request type |

Response Scoring

I. Response Rewrite Strategy Scoring:

Please refer to the table below on how to score the response strategy based on the prompt request type.

Generative/Mixed Prompts

| Prompt Toxicity | Risk Level | Decline | Disclaimer + Partial Engagement | Disclaimer + Full Engagement | Full Engagement |
|-----------------|------------|---------|---------------------------------|------------------------------|-----------------|
| Safe | All | 2 | 2 | 2 | 5 |
| Benign | All | 2 | 4 | 5 | 4 |
| Harmful | High Risk | 5 | 5 | 2 | 2 |
| Harmful | Low Risk | 2 | 5 | 2 | 2 |
| Jailbreak | All | 5 | 2 | 2 | 2 |

Extraction or Transformation Prompts

| Prompt Toxicity | Risk Level | Decline | Disclaimer + Partial Engagement | Disclaimer + Full Engagement | Full Engagement |
|-----------------|------------|---------|---------------------------------|------------------------------|-----------------|
| Safe | All | 2 | 2 | 2 | 5 |
| Benign | All | 2 | 2 | 2 | 5 |
| Harmful | All | 2 | 2 | 2 | 5 |

| | | | | | |
|-----------|-----|---|---|---|---|
| Jailbreak | All | 5 | 2 | 2 | 2 |
|-----------|-----|---|---|---|---|

II Contributor Response Scoring:

| Criteria | 1-2 (Fail) | 3 (Okay) | 4-5 (Good/ Perfect) |
|-----------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Rewrite Verbosity | <i>No failing due to this criteria</i> | [Verbosity Issues] <ul style="list-style-type: none"> - 1 or more responses answer the prompt but includes some key points that are irrelevant. - May repeat or rephrase the same idea more than once. - The length is fine, but it could be shorter. | <ul style="list-style-type: none"> - Claims include supporting evidence - Each paragraph or sentence adds value without unnecessary repetition. - Fits length required to answer prompt. |
| Rewrite Writing Style & Tone | [Major Writing Style & Tone Issues] <ul style="list-style-type: none"> • Spelling/Grammar: 5+ major spelling, grammatical, and punctuation errors that hinder readability • Fluency: The writing is clearly & objectively non-fluent <i>to the extent</i> that it sounds like a non-native language speaker • Tone: Response has an objectively disrespectful tone (i.e., teasing, mockery, foul language, negative expressions) or preachiness. • Text Formatting: Formatting of text is so unclear that it hinders readability • Phone Numbers: Response includes phone numbers | [Minor Writing Style & Tone Issues] <ul style="list-style-type: none"> • Spelling/Grammar: 1 or more responses have a few spelling, grammatical, and punctuation errors that minimally impact readability • Fluency: The writing is mostly fluent, but contains minor instances of non-native language • Tone: Marked by a somewhat disrespectful tone, mostly avoiding disrespect, teasing, mockery, foul language, any negative expression, or preachiness. • Text Formatting: Lists are used when appropriate, including bolding where relevant. Key details may not be bolded • First-Person Pronouns: | [No issues] <ul style="list-style-type: none"> • Spelling/Grammar: All responses have no spelling, grammatical, and punctuation errors • Fluency: The writing is perfectly fluent in the eyes of a native language speaker • Respectful: Marked by a respectful tone, avoiding disrespect, teasing, mockery, foul language, any negative expression, or preachiness. • Text Formatting: Lists are used when appropriate, including bolding where relevant. • First-Person Pronouns: Response uses zero first-person pronouns |

| | | | |
|-----------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| | | Response includes use of first-person pronouns | |
| Rewrite Truthfulness | [Major Truthfulness Issues] Includes significant inaccuracies which undermine the accuracy of the response. Common major inaccuracies include: <ul style="list-style-type: none"> • Unverifiable Claims: Claims that cannot be substantiated through credible evidence or reasonable validation • Fabricated Information: Citing nonexistent studies and sources, or fabricating details such as people, organizations, events • Contradiction of Established Knowledge: Contains blatantly incorrect or fabricated information about documented facts | [Minor Truthfulness Issues] Includes inaccuracies which affect the quality or precision of the response but are unlikely to produce disinformation, such as: <ul style="list-style-type: none"> • Imprecise Claims: Generalizations or overly broad statements • Omission of Details: Leaving out information that would improve the completeness of the response. | [No Issues] All responses are fully accurate, and meets all the below criteria: <ul style="list-style-type: none"> • All facts inside the response (such as definitions, numbers, dates, etc.) are completely accurate. • Response reflects majority / domain experts' opinions, or states clearly when there is no unique answer. • Clearly asserts that any false claims in the prompt are NOT true, rather than leaving room for ambiguity. • In cases where the user asks to summarize or rewrite a text segment, the response does NOT make up, or mention, details that were not part of the original text segment. |
| Rewrite Helpfulness | [Major Helpfulness Issues] <ul style="list-style-type: none"> • 1 or more responses could have clearly better fulfilled the user's request while still remaining concise and in-policy. • The response was a refusal, but clearly could have been a partial or full engagement | [Minor Helpfulness Issues] <ul style="list-style-type: none"> • 1 or more responses are appropriate but could probably have better fulfilled the user's request while still remaining safe, concise, and truthful. • The response was a refusal, but could have been more appropriate as a partial engagement • The response was a partial engagement but could have had a higher engagement | [No Issues] <ul style="list-style-type: none"> • There are no obvious ways any of the responses could have better fulfilled the user's request while still remaining safe, concise, and truthful. • The response's engagement level (refusal, partial engagement, full engagement) is appropriate. |

III. Response Labels Scoring:

Please refer to the table below on how to score the responses

| Criteria | 1-2 (Fail) | 3 (Okay) | 4-5 (Good/ Perfect) |
|--------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------|
| Rewrite Strategy NER Labels | [Major Rewrite Strategy Label Issues] 2 responses in a multi-turn response is clearly incorrectly labeled in terms of rewrite strategy or a single turn has been incorrectly labeled | [Minor Rewrite Strategy Label Issues] 1 response in a multi-turn response is clearly incorrectly labeled in terms of rewrite strategy or a single turn has been incorrectly labeled | [No issues] - All responses are mostly correctly labeled in terms of rewrite |