COURSE OF JANUARY 04, 2025

Escala de tiempo

Descripción generada automáticamente

Your task is to evaluate each response based on various quality dimensions.

This project has different work streams and quality dimensions tailored to their specific nature.

You will be presented with:

* User Prompt
* Possible context (e.g., active files or context sources)
* Two model responses

📣 **Important Note:**

Before starting any work on a workstream, make sure to "**Review the Instructions"** section at the beginning. This step is crucial for understanding the guidelines and requirements!

**How to Get Started**

You must access the reference **files/documents** for each task.

* Ensure you are signed into your browser and Google account with the email used for Remotasks/Outlier.
* If you cannot access the reference files, **stop tasking** and request access from your QM immediately.
* ***NOTE***: Attachments will appear in the response part most of the time but can also appear in the prompt part.

***Ground Rules:***

* All comments/justifications that you write must be in English so the project team can understand them.
* Evaluate responses based on specified task dimensions and descriptions **ONLY.**
* Consider the user's perspectives and what they are looking for before reviewing the answers.
* Ensure final Side-by-Side ratings align with individual response ratings. Ex: if Response A does better on groundedness, completeness, understanding, etc. it should also be selected with higher preference in the SxS rating.

**📊How to Assess the Prompt**

**Is the prompt ratable❓**

A prompt is **ratable** and you will be marked YES when:

1. The prompt is in the target language (your native language that you are tasking in) or in English. **📣Remember! English Prompts are also OK!**📣
2. The prompt is understandable - you understand what the prompt is asking you to do. If the prompt is nonsense or incomplete, mark the prompt as not rateable.

📍 Mid-conversation prompts are ratable.

📍 For tasks that include email responses ONLY - A prompt can simply include the content of the email that the user wants to send. For example: "I would like a refund within 2 weeks." Even though no explicit action is specified, it is clear that the user wants to generate an email with this content. *A prompt like this will be considered rateable*.

**How to Evaluate Individual Responses**

Rate each Model Response **individually** based on the following:

* Specified dimensions
* Ex:
* **Next to each option in a dimension, there is an information icon Aplicación, Icono

  Descripción generada automáticamente con confianza mediathat you can hover over to see its description.**
* Provide an overall rating with justification
* Make sure the overall rating is **aligned** with your prior ratings + justification. Explain why you chose specific dimensional ratings.
* When a response makes a factual statement, confirm that the statement is true via outside research.
* Ex: if the response claims that the capital of Colombia is Bogotá, Google this to confirm it is true.
* Ex: if the response is summarizing claims about a document, verify the claims inside the actual document.
* First, rate each response as a standalone.

**🔎Dimensions to Evaluate the Response**

***IMPORTANT NOTE:*** *REFER TO YOUR PROJECT'S SPECIFIC INSTRUCTIONS SINCE EDGE CASES MIGHT DIFFER ACROSS THE PROJECTS BUT THE BELOW IS A GENERAL OUTLINE OF WHAT TO EXPECT.* ***SPECIFIC PROJECT INSTRUCTIONS TAKE PRECEDENCE OVER THE COURSE IF THEY ARE IN ANY WAY NOT ALIGNED***

📍 **Ratability**

* Responses **must** be in the target language: Even though the **prompts** can be in English, **the responses cannot be in English**.
* The responses and attachments must be understandable enough to assess if they successfully answer the user's prompt.
* No Personally Identifiable Information (PII): (names and surnames are not considered PII). Examples of PII are:
* Social Security number, physical address, email address, driver's license number, bank account number, passport number, Date of birth
* Non-Harmful: The response should not be harmful.

🔖 Completeness Dimension

Completeness asks you to evaluate if the response addresses all parts of the user’s request. This dimension is not related to the context files.

﻿For example, if the prompt asks for 3 things, we must ensure that the response returns the 3 things that were asked for.

Example prompt:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Good Completeness response:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

**🔖 Understanding Dimension**

This dimension assesses how well YOU understand this response.

For example, if the prompt asks about summarizing an attachment, the response must be well presented so that YOU can understand the information in the response about what is asked for in the prompt.

Despite the truthfulness of the response, the information is perfectly understandable.

**🔖 Truthfulness Dimension**

***What are Factual Claims?***

Factual claims are information expected to be correct and verifiable through research. Focus on objective claims that can be confirmed or refuted with specific information. Consider subjective claims accurate unless strong evidence suggests otherwise.

Not all responses will have factual claims. If none are present or if the prompt asks for creative ideas, mark this question as “N/A.”

⚠️ Verify claims using tools like Google to research!

**🔖 Groundedness Dimension**

Groundedness rates accuracy of the response based on information **WITHIN the context of the files/emails**.

﻿The difference between the Groundedness and Truthfulness Dimension is that Groundedness is based on the information we can find on the internet, Groundedness is based on the information we can find within the attachments.

**Prompt example:**

“When did Argentina gain independence?”

**Bad Groundedness example:**

Aplicación

Descripción generada automáticamente con confianza media

**Document:**

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

In this case, the model is wrong, as the document clearly states that Argentina's independence was in 1816.

**⚖️ How to Compare and Rank Responses Side-by-Side**

In this part, you will compare the 2 responses on a scale of 1-7 to decide which response is better and provide a justification.

🔴 The side-by-side justification should be 2-3 sentences and highlight the most important factors influencing your preference.

🔴 A few guidelines for what a ‘better’ response looks like:

* “Better’ responses are *grounded* and *truthful*, meaning that the presented facts are based on the context provided and verifiable using external sources.
* If multiple responses are similarly (in)correct, consider which response is most likely to be *helpful*, meaning the response matches the prompt and provides a useful starting point for the user.

! Please make sure your final rating is consistent with the ratings you gave for each response individually.

Example: You rated

* Response 1 'Okay'
* Response 2 'Amazing'

❌ **You should not then rate Response 1 as better than Response 2.**

\* If your Response comparison rating does not match your individual ratings for each Response, you will see a red 'Linter' error.

**GROUNDEDNESS**

Remember:

* Groundedness rates accuracy of the response with respect to information WITHIN the context files/emails.
* If the response is accurate GIVEN the context documents, it IS grounded
* If the response is NOT accurate GIVEN the context documents, it is NOT grounded
* Use the context files to evaluate whether a response is grounded.

If you are an ATTEMPTER:

* Ensure that you are assessing groundedness accurately and correctly!

If you are a REVIEWER:

* Ensure that all attempts have rated groundedness correctly. If not, you MUST correct the attempt

I will evaluate groundedness correctly

Dear contributors,

We’re about to release tasks attempts that contain the following prompt: **Summarize this document**.

**This prompt should be marked as RATABLE since the focus of this batch is to taste how good the model is in summarization (focus is not in prompt this time), attached files are in expected language**

It’s essential you don’t rate them as: *not in the target language*, same goes for reviewers, if they find a case like this they should fix it.

Side by side scoring guide:

1 if both responses are rateable, select which response you think is better (also consider how much better). Select 1,2,3 to favor response 1, 4 for no preference, 5,6,7 to favor response 2. Make sure to weigh groundedness, understanding and completeness heavily over other evaluation dimensions, ensure your overall response scores (horrible, pretty bad, amazing, etc) for each response is consistent with the side by side score, just select no preference 4 if both responses are horrible-pretty bad, do not spend time on deciding which one is slightly worse

2 if bot responses are unratable, strictly select no preference 5 in the side-by-side score

3 if exactly one response is ratable, if the ratable response is pretty bad or horrible select no preference 4. If the ratable response is okay or better, always select a value that favors the ratable response. For example if response 2 is the rateable response, depending on how good response 2 is make a selection between 5, 6 or 7, if its amazing is 7, if is pretty good is 6, if is okay is a 5

1 response 1 is much better

2 response 1 is better

3 response 1 is slightly better

4 no preference

5 response 2 is slighty better

6 response 2 is better

7 response 2 is much better

Justifications section

**Justifications**

**⚠️Remember - all justifications need to be in English 🇺🇸**

**🧐 Why Do We Need Justifications**

**What is Justification?**

This is the rationale you provide for your evaluation of a particular response to a prompt and/or to your selection of one response over another.

**Why are good justifications important?**

Justifications allow reviewers, the project team, and the client to better understand your thinking and preferences so that we can provide accurate feedback, useful and targeted training, and improve model alignment with human preferences (e.g., values, satisfaction).

Justifications, in short, explain why you picked one response over the other. Crucially, you must write a very beautiful, holistic justification.

In the TNFM project there are multiple questions where justifications are required! The justifications guidelines apply to ALL JUSTIFICATIONS BOXES!

**😡 Let's Chat about Lousy Justifications**

Critiques must be specific and include supporting claims with the deficits of the response.

Critiques must be coherent without contradictions, must be correctly in grammar, and provides supporting claims.

All the justifications must be in English.

**A good justification includes your reasoning, specific examples from the responses, and how they influenced your ratings.**

🧑‍🏫Elements of a good justification

* Aim for at least **4-5 sentences**
* Lead with what **the user intent** is (What does the user want to achieve?)
* Includes **key differences** between the responses (How is summary/recap A different to summary/recap B)
* Highlight these factors with reasonably detailed evidence (use examples from the text!)
* INCLUDE appropriate textual **evidence to support your verdict**.
* **ALIGN** it with your verdict (i.e. rating it “much better” requires stronger evidence) **MAKE SURE THERE ARE NO CONTRADICTIONS BETWEEN JUSTIFICATIONS AND RATINGS.**
* Show your research & your work, based on the dimensions:
* If an issue (eg grammar or something that doesn’t sound like native) is identified for writing quality → share it in your comment
* If a summary/recap is not grounded on the transcripts explain exactly what makes it not **GROUNDED** - give examples.
* If a recap/summary is not **COMPLETE,** explain what should have been included!
* If a response has other issues, explain what are these!
* **Justifications without evidence and supporting claims from the responses or your research will not be counted as good.**

**🧱Elements of good justifications:**

* **User Intent**: The justification should include a comment that demonstrates an understanding of what the user is looking for with the prompt
* **Conclusion**: We need to determine which response is better and why. The comment should clearly explain the reasoning behind this decision.
* “@Response A is better than @Response B ...”
* **Supporting Claims**: The justification needs to have key supporting points to defend its conclusion.
* “...because @Response B has a hallucination issue ... "
* **Specific Evidence**: The precise examples or evidence in the text used to support each supporting point.
* “..Response B has a groundedness issue in the second paragraph when its talking about”

﻿

* **Analysis**: The explanation(s) of how the evidence defends the supporting claim
* Summarize how your evidence supports your preference

🤩 Guiding Principles for Good Justification

Focus on ultimately why you selected your preference (i.e., which response best aligned/answered the user's core intents)

Be specific about what the response(s) might have missed (i.e., if you say it didn't follow what the user asked for, say specifically what it didn't follow; if you say it had major factual inaccuracies, say precisely what it got wrong)

**Edge Cases / Issues with files 📄**

Sometimes the files provided in the task are not in the ideal shape or form, which might be confusing, but don’t worry! **Here is exactly what to do** in these situations!

✏️In the case where **attachments/files** are **extremely long (40 pages or more)**

* Proceed with marking:
* The response as **Unratable**

✏️In the case where **attachments/files** are **either** not present, not relevant to the prompt or not in the target language.

*Scenario 1:* Response is a punt:

**Example**:

**Prompt**: Summarize the provided renter's agreement?

**Response**: There’s not enough information

**Attachments**: *Blank, totally irrelevant, or in the wrong language*

Proceed with marking:

* The response as **Ratable**
* Completeness as **Complete**
* Groundedness as **Completely** **Grounded**
* **Evaluate all other questions regularly**

*Scenario 2:*The Prompt can be answered without any files/context and the response answers the prompt (not a punt)

**Example**:

**Prompt**: What are the colors of the rainbow?

**Response**: The 7 colors are red, orange...

**Attachments**: *Blank, totally irrelevant, or in the wrong language*

Proceed with marking:

* The response as **Ratable**
* Completeness as **Complete**
* Groundedness as **N/A**
* **Evaluate all other questions regularly**

*Scenario 3:* The prompt asks about a reference attachment and the response answers directly (not a punt)

**Example**:

**Prompt**: What does the second section of the contract state?

**Response**: The second section of the contract is about employee benefits...

**Attachments**: *Blank, totally irrelevant, or in the wrong language*

Proceed with marking:

* The response as **Ratable**
* Completeness as **Incomplete**
* Groundedness as **Ungrounded**
* **Evaluate all other questions regularly**

**Workflow Basics 🔄**

**Understanding the proper workflow ensures consistent and thorough evaluations.**

* **Ensure you have a Gmail account as your primary or secondary email in Outlier settings:** Essential first step for accessing all necessary resources
* **Make sure you're signed into this Gmail account in your browser!** This is a MUST have to task on this project!
* **Access reference files:** Gather all the context and materials needed for evaluation.

*Pro Tip: Always complete steps in order - skipping steps can lead to incomplete evaluations.*

*If you cannot see reference files while tasking, reach out to your QM.*

**Step [1]: Prompt Assessment 📝**

**Prompt Assessment 📝**

A good prompt evaluation sets the foundation for accurate response assessment.

**Is the prompt ratable?**

* ✅ **Target Language Compliance:** Ensures communication clarity and consistency
* ✅ **Understandability:** Must be clear enough for both human and AI comprehension
* ✅ **Explicit Request:** Should clearly state what is being asked for
* ❌ **Avoid Ambiguity:** Remove any unclear or vague elements

Pro Tip: If you're unsure what the prompt is asking for, it's probably not ratable.

**Step [2]: Response Evaluation Basics 🎯**

**Response Evaluation Basics 🎯**

Every response must meet these fundamental criteria before detailed evaluation.

**Key Criteria:**

* **Language Compliance:** Must match the required target language
* **Understandability:** Clear and coherent communication
* **Privacy Protection:** No personal identifiable information (except names/surnames)
* **Safety:** Content must be non-harmful and appropriate
* **Completeness:** All required files and attachments are present
* **Special Cases:** Proper handling of "punt" responses

***Pro Tip:*** *Think of this as your initial checklist before diving deeper. If you asses that the response is a* ***PUNT,*** *follow the* [*guidelines here.*](https://docs.google.com/document/d/e/2PACX-1vSIqPtUC0GUGEHc7ndh-hwH7VE-AZ8EECwtVGHX2UXRhAmsNjBTbfbO9s42H7AvjYCmtt_P-EsAmGbd/pub#h.gdvc64at4ybo)

*We will dive deeper into each of the key criteria in the following slides -->>>*

**Step [3]: Response Rating Rubrics**

**Attachment Visibility:** Do you have access to view these reference items?﻿

***Note: If you do not have access to the attachments, please skip the task and talk to your QM immediately.***

**Is the response rateable?** Yes/No

* The response is in the target language **(English is not allowed like in the prompt section)**.
* The response and attachment are understandable enough to assess if they complete the prompt.
* If there are any issues with attachments, refer to the [Issues with Files](https://docs.google.com/document/d/1Pfm_t1pHtYuCatTy-ZqN9HC0fiDseLa2MkmYS2xRjIU/edit?tab=t.0#heading=h.58erdglnwlg0) instructions in the Appendix.
* The response does not include personally identifiable information (PII). (PII does NOT include names and surnames).
* The response is not harmful.
* The response includes a file (unless the file is not relevant, [details](https://docs.google.com/document/d/1Pfm_t1pHtYuCatTy-ZqN9HC0fiDseLa2MkmYS2xRjIU/edit?tab=t.0#heading=h.58erdglnwlg0)).
* If the response is a **punt**, please follow the guidelines [here](https://docs.google.com/document/d/e/2PACX-1vSIqPtUC0GUGEHc7ndh-hwH7VE-AZ8EECwtVGHX2UXRhAmsNjBTbfbO9s42H7AvjYCmtt_P-EsAmGbd/pub#h.gdvc64at4ybo).

*NOTE: When the prompt is* ***not asking to retrieve a file****(i.e. a summarization prompt) and the model response is just the name of a file (in English), the response should be Ratable, Incomplete, not grounded, and understandable*

**File Quality:** How good is the quality of the attachments or files? Are any of these issues present? Select a maximum of 3 choices.

**Completeness**: How well does the response fulfill the user's request?

Complete if the response follows all instructions and fully delivers on the request of the prompt. For summary projects a response is complete if the response is a summary covering all the important topics or main ideas that are essential to understanding the content of the files

Partially complete if the response addresses most of the instructions or goals of the prompt, but misses or misinterprets some of the key points

Incomplete if the response misses key instructions or goals of the prompt, rendering it not helpful to the user

**Understanding**: How easy and clear it is for the user to understand the model response based on the prompt provided

Understandable if the response is easy to understand, well written, coherently organized, and not repetitive

Partially understandable if the response can be understood, but has minor issues with writing quality, organization or repetition

Not understandable if the response is barely intelligible, very confusing, or so poorly organized that it was difficult to understand

**Groundedness**: How correct and consistent is the model response based on the user's prompt, using the context reference(s) if any

Completely grounded if all information in the response is correct and consistent, based on the context references if any

Reasonable grounded if important information in the response is correct or consistent, based on the context references. However, the response includes minor inconsistencies or errors that are not essential to answering the prompt

Not grounded if at least one piece of important information in the slide content is incorrect or inconsistent, based on the context references

N/A if there are no retrieved resources for this response

**Retrieved sources correctness [ONLY FOR Corpus Q&A]: How credible and correct are the sources retrieved by the model response based on the user's prompt**

High quality sources if the model documentrs has high quality sources

Medium quality sources if the model documents sources are of medium quality

Poor quality sources if the models documents sources are of poor quality

**Overall Response Rating and Justification:** How well does the response answer the user's prompt overall and why? This should be an overall assessment of the response building upon the rubrics tackled above and adding other elements that are important for a response to be stellar vs subpar vs average.

Amazing, this response really delivered on the prompt. It was helpful, complete, factual, and creative where appropriate. Media: if the response included non-text output (e.g. images or video), the output was so aesthetically appealing that it seemed like it was created by a professional. You would definitely use this LLM again and would recommend it to others.

Pretty good: this response was good overall. It was mostly helpful, and creative where appropriate. Media: If the response included non-text output (e.g. images or video), it was aesthetically appealing and did not contain any flaws/distortions. You would probably use this LLM again.

Okay: This response was fine, though it didn’t leave much of an impact. It may have had minor issues following instructions, had akward formatting, or provided unoriginal ideas. Media: if the response included non-text output (e.g. images or video), the output did not have major flaws/distortions but it wasn’t particularly impressive. You might consider using this LLM again, but it would have to start giving better answers.

Pretty bad: This response would require a significant amount of manual editing to be useful. It had major issues with following instructions, including false information, incomprehensibility, and missing opportunities to be creative where appropriate. Media: If the response included non-text output (e.g., images or video), that output was not aesthetically appealing, and it may have had flaws/distortions. You would probably avoid using this LLM.

Horrible: This response really missed the mark. It did not follow instructions, included false information, was hard to comprehend, or missed opportunities to be creative where appropriate. Media: If the response included non-text output (e.g. images or video), that output was not aesthetically appealing at all had major flaws/distortions. You would actively avoid using this LLM in the future and would tell others not to use it.

***Note:*** Only rate responses as **"Amazing"** if they meet **all the specified criteria**.

**Step [4]: Side-by-Side Comparison and Justifications**

**Key Points:**

1. Compare the groundedness between responses
2. Assess relative "helpfulness"
3. Check "prompt adherence" consistently
4. Use the Likert (1-7) scale for comparison
5. Maintain rating consistency with individual assessments
6. Provide a clear and strong justification referencing the two responses as to why you have chosen the selected rating on the Likert.

***Pro Tip: Always refer back to your individual ratings to ensure consistency.***

While tasking, please avoid the following common errors at all times: inconsistencies between individual ratings and overall rankings, incorrect assessments of language eligibility, inadequate justifications for ratings, and overlooking key details in the content provided.

* **Ensure Consistency Between Individual and Overall Ratings**: When providing individual ratings for responses, make sure the overall rating aligns with the details. If Response 1 is rated better than Response 2 in individual criteria, this should be reflected in the side-by-side comparison and overall ranking.
* **Language Appropriateness**: Always assess whether a response is in the expected language before rating. If a response is in the wrong language, mark it as "Not Ratable". Remember that prompts in English are OK but the responses should be in the assigned languages.
* **Thoroughly Read and Ground Justifications**: Review the content of emails or attachments carefully. Ensure your ratings and justifications are based on accurate interpretations of the information provided, paying attention to groundedness and completeness.
* **Match Justifications to Ratings**: Justifications must clearly explain the reasons for each rating. Avoid short, identical justifications for different responses, and ensure that explanations align with the overall rating and dimensions such as completeness and helpfulness.

Common Mistake: Quality Inconsistency

When providing feedback on responses, it's crucial to ensure consistency in rating and commenting across both Responses 1 and 2.

It has been observed that Response 1 may receive a higher quality rating with detailed comments, while Response 2 is rated poorly and with little explanation.

This discrepancy can lead to misunderstandings about the evaluation process and hinder improvement efforts.

To maintain fairness and clarity, contributors should provide thorough assessments for each response, regardless of their satisfaction level. This practice ensures that all feedback is constructive and valuable

**Truthfulness**

When evaluating the truthfulness of chatbot responses, it is essential to ensure that the information provided is accurate and reliable.

It has been observed that a response may sometimes present misleading or incorrect details, which can negatively impact user trust and create confusion. Maintaining truthfulness is crucial because users depend on the chatbot for reliable information, and any deviation from factual accuracy can undermine the integrity of the entire interaction.

To enhance the value of the responses, the chatbot should consistently verify and convey accurate content, providing well-supported and verifiable information. This approach ensures that users can trust the system and find the answers they receive genuinely helpful and informative.

Recommendations

Oftentimes the chatbot will not make claims that need to be verified due to the nature of the projects, but when it does, it is very important to make sure those are factually correct. A quick search online can show if the response is truthful or not.

**Common Mistake: Justification's Quality**

When providing justifications, it's essential to ensure clarity and grammatical accuracy. Many submissions continue to exhibit grammar issues, which can obscure the intended message and reduce the effectiveness of the justification. Additionally, justifications often lack clear conclusions, leaving readers without a solid understanding of the rationale behind the rating.

Recommendations:

To improve the quality of feedback, contributors should proofread their justifications for grammatical errors and ensure that each justification includes a concise conclusion that ties together their points. This approach will enhance communication and provide clearer insights into the evaluation process.