Multi-Challenge V2 Conversation Quality Review, Rubric Editing, & Evaluation Instruction

Last Updated: February 12, 2026

Quick Start

Your Mission: Review multi-turn conversations, evaluate pre-provided rubric criteria against 3 model responses, edit rubrics when necessary, and rate each rubric on quality dimensions.

Before You Begin:

Read this guide completely
Understand the 4 failure axes (testing axes)
Know the conversation quality criteria
Understand how to evaluate and edit rubric criteria
Know how to rate rubric quality (Importance, Specificity, Atomicity, Verifiability, Difficulty)
Have project access ready
Know who to contact with questions

Project Access

Labelbox Project: https://app.labelbox.com/projects/cmlimx4h00ahb07tpbfc9a040
Project Name: Multi-Challenge-V2 (094) Feb-26
Need Help?

Questions: Contact your project manager
Technical Issues: Labelbox support
Clarifications: Use team communication channel

Table of Contents

Understanding the Data
The 4 Failure Axes (Testing Axes)
Conversation Quality Criteria
Review & Validation Process
Evaluating Rubric Criteria
Rating Rubric Quality
Labelbox Fields Guide
FAQ
Quick Reference Cheat Sheet

Understanding the Data

Each task contains a multi-turn conversation between a User and AI assistants:

Prior turns: The User exchanges messages with a single Assistant (context turns).
Last turn: The User's final message receives 3 separate model responses side by side:

Assistant 1
Assistant 2
Assistant 3

Each task also comes with:

A Testing Axis label (pre-filled, read-only) that tells you what type of failure the conversation was designed to test.

Pre-provided rubric criteria attached to the last turn. These are specific, checkable statements about what the response should or should not contain. You will rate each model response against these rubrics, edit the rubric text if needed, and evaluate the quality of the rubric itself.

The 4 Failure Axes (Testing Axes)
Each conversation is designed to test one of these failure types. The Testing Axis field (read-only) tells you which one.
1. Instruction Retention
Definition: The assistant forgets an explicit instruction from the first turn that is stated to be maintained throughout the entire conversation.
Look for:

User gives explicit command in Turn 1
Assistant does not follow that instruction in a later turn

Example:
User (Turn 1): "Please use only metric units in all your responses."
Assistant (Turn 1): "Got it, I'll use metric."
User (Turn 2): "How tall is the Eiffel Tower?"
Assistant (Turn 2): "The Eiffel Tower is 1,083 feet tall." [FAIL]
Key Points:

Instruction must be EXPLICIT it applies for the whole conversation ("please always…", "make sure to always…")
Must be from the turn 1 user instruction

2. Inference Memory
Definition: The assistant forgets, misremembers, or misapplies information that is provided by the user.
Look for:

User shares personal info, preferences, or context
Later asks for recommendations or suggestions
Assistant fails to use that information provided by the user appropriately

Example:
User (Turn 1): "I'm vegetarian and allergic to nuts."
Assistant (Turn 1): "Noted! How can I help?"
User (Turn 2): "Suggest a healthy snack."
Assistant (Turn 2): "Try some honey-roasted almonds!" [FAIL]
Key Points:

User shared information
Assistant should have inferred how to use that info

3. Self-Coherence
Definition: The assistant is not coherent with something the assistant previously said. (The requirement is that it is a different assistant turn)
Look for:

Assistant makes a statement in one turn

Later contradicts that statement in another turn
Comparing assistant turn to assistant turn

Example:
User (Turn 1): "..."
Assistant (Turn 1): "The event starts at 3 PM on Saturday."
User (Turn 2): "Great, I'll be there."
Assistant (Turn 2): "..."
User (Turn 3): "What time was that again?"
Assistant (Turn 3): "The event is at 2 PM on Sunday." [FAIL]
Key Points:

Must be assistant contradicting ITSELF from a previous turn
Comparing two different assistant responses

4. Reliable Version Editing
Definition: The assistant fails to make a change that is requested by the user to a previous assistant turn or makes a change that was not requested. (This can be any previous assistant turn)
Look for:

User asks for specific changes to content
Assistant fails to apply the requested edits

Example:
Assistant (Turn 1): "Here's your grocery list: milk, eggs, bread, cheese."
User (Turn 2): "Remove cheese and add butter."
Assistant (Turn 2): "Updated list: milk, eggs, bread, cheese." [FAIL]
Key Points:

User explicitly requested an edit
Edit must be to previous content

Conversation Quality Criteria
Check These FIRST (~5 minutes)
The conversation should be rejected (via the quality check) if ANY of these are true:
1. USER Turn is CUT OFF

USER requested changes but didn't provide the draft
Conversation is incomplete or missing content
Example: "Please edit this: [nothing provided]"

2. User or Assistant REFUSES to Comply

User fails to make a valid request or Assistant declines to answer or help with valid request
Example: "I cannot help with that" when request is reasonable

3. Non-ENGLISH Language

Reject any conversation not in English
If the model provides single word or quick phrase translation from a language other than English into English, this is ACCEPTABLE!

4. Assistant CANNOT Perform Task
Examples:

Create diagrams, graphs, images
Calendar invites
System actions beyond text generation

5. Inappropriate Content

Personal or specific medical advice
Sexual content
Illegal content

6. Model References Itself by Name

Example: "I'm ChatGPT, an AI language model developed by OpenAI"
ASSISTANT should not identify itself by specific name

Review & Validation Process
Overview of Steps

Check the Testing Axis (read-only field, pre-filled)
Read the entire conversation carefully
Verify failure alignment with the Testing Axis
Check conversation quality (pass/fail)
Evaluate each rubric criterion against each model's response (meets_criterion/does_not_meet_criterion)
Edit rubric criteria if the text is unclear, incorrect, or could be improved
Rate rubric quality (Importance, Specificity, Atomicity, Verifiability, Difficulty)
Select the Best Response among the 3 assistants
Rate Overall Satisfaction for each assistant's last turn

STEP 1: Check the Testing Axis
The testing_axis field is pre-filled and read-only. It tells you the type of failure this conversation was designed to test (e.g., "Inference Memory", "Reliable Version Editing"). Use this information to guide your review of Assistant 1's behavior.

STEP 2: Read the Entire Conversation

Read from start to finish carefully
Pay attention to:

Every user instruction
How responses adhere to the user's preferences
Claims made by the assistant in prior turns
Changes across turns
The differences between the 3 model responses in the last turn

STEP 3: Verify Failure Alignment (failed_axis)

Question: Does Assistant 1's last turn failure reason align with the Testing Axis above?

Yes: Assistant 1's response in the last turn fails in the way described by the Testing Axis.

Then proceed to check conversation quality.

No: The failure does not match the Testing Axis (or there is no failure).

Provide a brief justification explaining why.
Since you chose "No," this row is not useful. Provide the justification and you may complete remaining fields quickly and submit.

STEP 4: Check Conversation Quality (passes_conversation_quality_checks)
This field appears only if you selected "Yes" for failed_axis.
Question: Does the conversation between User and Assistant 1 pass all conversation quality rubrics?

Yes: The conversation is well-formed and usable. Proceed to rubric evaluation.
No: The conversation has quality issues (see criteria above). Provide a detailed rejection reason.

Since you chose "No," this row is not useful. Provide the justification and you may complete remaining fields quickly and submit.

Evaluating Rubric Criteria
What Are Rubric Criteria?
Each task comes with pre-provided rubric criteria attached to the last turn. These are specific, checkable statements about what a model response should or should not do. Examples:

"The response should offer at least two distinct herb substitution options."
"The response should not use the verb 'directed' to describe the relationship with engineering and design teams."

How to Evaluate
For each rubric criterion, evaluate each model response separately:
Rating Options:
Rating
When to Use
meets_criterion
The model's response satisfies the rubric criterion
does_not_meet_criterion
The model's response fails to satisfy the rubric criterion
Optional: Explain Reason
Each rating has an optional nested text field: "Explain Reason". Use this to clarify your decision when:

The criterion is borderline or ambiguous
Your rating might seem surprising
You want to highlight specific evidence in the response

Editing Rubric Criteria

You can and should edit the rubric criterion text if:

The criterion is poorly worded or ambiguous

The criterion contains factual errors

The criterion is too broad and should be more specific

The criterion conflates multiple checks into one (not atomic)

However, you cannot delete rubrics. You can only edit and rate them.

Example Evaluation

Rubric: "The response should offer at least two distinct herb substitution options (not counting dried dill) for the user to choose from."

Model

Rating

Reason

Assistant 1

meets_criterion

Suggests parsley, tarragon, and chervil as alternatives.

Assistant 2

does_not_meet_criterion

Only mentions dried dill as a substitute.

Assistant 3

meets_criterion

Provides parsley and fennel fronds as alternatives.

Rating Rubric Quality

For each rubric criterion, you must also evaluate the quality of the rubric itself using these dimensions:

1. Rubric Importance

Question: How important is this criterion for evaluating the response quality?

Rating

Description

Essential

This criterion addresses a core requirement that directly impacts whether the response is useful and correct

Important

This criterion addresses a meaningful aspect of response quality, but not the most critical

Optional

This criterion is nice-to-have but not necessary for a good response

2. Specificity Score

Question: How specific is this criterion to the particular prompt/conversation?

Rating

Description

Not Specific to the Prompt

The criterion is generic and could apply to almost any conversation

Somewhat Specific to the Prompt

The criterion relates to the topic but doesn't reference specific details

Very Specific to the Prompt

The criterion directly references specific details, names, numbers, or context from the conversation

3. Atomicity Score

Question: Does this criterion test a single, focused behavior?
Rating
Description
Not Atomic (Multiple Behaviors Mixed)
The criterion bundles several checks (e.g., content + formatting) into one
Partially Atomic
Mostly focused on one behavior, but still has small extra conditions or side-requirements mixed in
Fully Atomic (Single Behavior Only)
The criterion clearly targets one behavior or property

4. Verifiability Score
Question: Can the criterion be objectively checked against the response?
Rating
Description
Not Verifiable
Vague or subjective (e.g., "the response should be good")
Partially Verifiable
Some concrete anchors but still requires judgment
Concrete and Checkable
Can be definitively verified as met or not met

5. Difficulty Score
Question: How challenging is this criterion for a model to satisfy?
Rating
Description
Too Easy / Trivial
The criterion can be satisfied by almost any reasonable response
Appropriately Challenging
The criterion meaningfully separates good responses from mediocre ones
Too Hard / Unreasonably Strict
The criterion demands edge-case or expert-level behavior

Labelbox Fields Guide (STEP BY STEP)
Complete Field-by-Field Instructions

1. Testing Axis (Read-Only)
Field: testing_axis
Type: Text (read-only, pre-filled)
Action: Do NOT change this field. It tells you the expected failure type for this conversation.

2. Failed Axis
Field: failed_axis
Type: Radio button (required)
Options: Yes | No
Question: Does the Assistant 1's last turn failure reason align with the Testing Axis above?
Choose "Yes" when:

Assistant 1's last response clearly fails in the way described by the Testing Axis
The failure is genuine (not caused by conflicting user instructions)
The failure originates from earlier conversation history (not just the last user turn, except for Reliable Version Editing)

Choose "No" when:

The failure doesn't match the testing axis
There is no failure at all
The failure is caused by the user (conflicting instructions, hallucination)

If "Yes" → Complete Passes Conversation Quality Checks (see below)
If "No" → Complete Does Not Fail Designed Axis justification text. Since this row is not useful, you may complete remaining fields quickly and submit.

3. Passes Conversation Quality Checks (nested under "Yes")
Field: passes_conversation_quality_checks
Type: Radio button (required, shown when failed_axis = Yes)
Options: Yes | No
Question: Does the conversation between User and Assistant 1 pass all conversation quality rubrics?
Choose "Yes" when:

No auto-reject criteria are triggered (see Conversation Quality Criteria above)
The conversation is realistic and coherent
The user has not hallucinated or caused the model error

Choose "No" when:

Any auto-reject criterion is triggered
Provide a detailed justification in other_quality_reject_reason
Since this row is not useful, you may complete remaining fields quickly and submit.

4. Best Response: Selection
Field: Best Response: Selection
Type: Message single-selection (required)
Action: Select the best model response among the 3 assistants in the last turn.
How to choose:

Consider accuracy, helpfulness, adherence to user instructions, and overall quality
Consider how well each response handles the conversation context from prior turns
If all responses are equally good (or bad), pick the one that best addresses the user's needs

5. Overall Satisfaction for Assistant 1's Last Turn
Field: Overall Satisfaction for Assistant 1's Last Turn
Type: Radio button (required)
Options: Amazing | Pretty Good | Okay | Pretty Bad | Horrible
Rate the overall quality of Assistant 1's last response:
Rating
Description
Amazing
Exceptionally helpful, accurate, well-structured, and perfectly addresses the user's needs
Pretty Good
Solid response with minor room for improvement
Okay
Adequate but has noticeable shortcomings
Pretty Bad

Significant issues that reduce usefulness
Horrible
Completely fails to address the user's needs or contains major errors

6. Overall Satisfaction for Assistant 2's Last Turn
Same as above, but for Assistant 2.

7. Overall Satisfaction for Assistant 3's Last Turn
Same as above, but for Assistant 3.

8. Rubric Criteria Evaluation (Per Rubric)
For each pre-provided rubric criterion on the last turn:
a) Rate each model response: meets_criterion or does_not_meet_criterion
b) (Optional) Explain Reason: Provide justification for your rating
c) Edit the rubric text if it is unclear, incorrect, or needs improvement

9. Rubric Quality Ratings (Per Rubric)
For each rubric criterion, rate these quality dimensions:

Rubric Importance: Essential | Important | Optional
Specificity Score: Not Specific to the Prompt | Somewhat Specific | Very Specific
Atomicity Score: Not Atomic | Partially Atomic | Fully Atomic
Verifiability Score: Not Verifiable | Partially Verifiable | Concrete and Checkable
Difficulty Score: Too Easy | Appropriately Challenging | Too Hard

See Rating Rubric Quality above for detailed descriptions of each option.

Specific Rejection Rules
The error must occur in the last turn, but it must be caused by the earlier conversation history. It cannot be caused by the user on the last turn. Such errors do not count as valid failures (cannot have failed_axis = Yes), except for Reliable Version Editing tasks.
1. Self-Coherence Error ONLY Within Last Assistant Turn

The error must originate from a previous ASSISTANT response
If contradiction exists entirely within one response, this is a logic error, not a memory failure
Another error must exist in the final turn for this to have failed_axis = Yes

2. Self-Coherence About Factual Error Later Corrected

Assistant gave wrong info, then corrected it
Self-correction is acceptable behavior
Another error must be present for this to have failed_axis = Yes

3. Inference Memory Error from LAST USER Turn

Constraint must be from earlier turns, not mentioned in the most recent
Example: User mentions diet in turn 1 → OK
Example: User mentions diet in last turn → Not a valid MultiChallenge failure

4. Model ONLY Fails on Instruction from LAST USER Turn (EXCEPT Reliable Version Editing)

Memory failure FOR INSTRUCTION RETENTION MUST STEM FROM TURN 1
If instruction is brand new in last turn → Not a valid MultiChallenge failure

5. User Asks for "a" or "one" Recommendation, Model Gives Multiple

This is ACCEPTABLE, NOT a failure
Model being helpful by providing options
Don't reject for this

6. Abstract Request for "Brief/Short" and Response is ≤ 250 Words

Use 250 words as guideline
≤ 250 words = acceptable as "brief" or "short"
Don't reject if within this threshold

FAQ
Q: What if there are multiple failures across the 3 model responses?
A: Focus on evaluating each rubric criterion independently against each model response. The rubrics are designed to capture specific aspects of quality.
Q: How long should this take?
A: Typically 15-25 minutes per task. Most time should be spent on rubric evaluation.
Q: What if a rubric criterion doesn't make sense for the conversation?
A: Edit the rubric text to make it more appropriate, or rate it as "Optional" importance and note the issue in the Rubric Score reason.
Q: Should I edit every rubric?
A: No. Only edit rubrics that are unclear, incorrect, or need improvement. If a rubric is well-written, leave it as-is and just rate it.
Q: What if all 3 responses are equally good?
A: Still select one as "Best Response." Use your best judgment on which is marginally better. Rate all three with similar Overall Satisfaction scores.
Q: What if the failed_axis is "No"?
A: Provide your justification, then you may complete the remaining fields quickly and submit. This row is not useful for the project.
Q: How do I know if a rubric is "Atomic"?
A: Check if it tests ONE thing. "The response should list 3 herbs" is atomic. "The response should list 3 herbs AND explain their flavor profiles AND mention cooking times" is not atomic.
Q: Failure only from the last user turn (not a Reliable Version Editing failure)?
A: Select "No" for failed_axis. The failure must be related to the conversation history from earlier turns.
Q: Self-Coherence error within a single turn?
A: That's a contradiction within one response, not a multi-turn failure. There must be another failure that originates from the conversation history.
Q: Assistant corrected its own error?
A: Self-correction is acceptable behavior. A different error must be present for the failure axis to be valid.

Quick Reference Cheat Sheet
The 4 Failure Axes
Axis
Quick Test
Instruction Retention
User said "do X for our whole conversation", assistant later doesn't
Inference Memory
User shared info, assistant ignores it
Self-Coherence

Assistant said X, later says not-X or does not remember saying X
Reliable Version Editing
User: "change X to Y in Z", assistant keeps X or doesn't remember Z correctly
Workflow Summary

```
1. Read Testing Axis (read-only)
2. Read entire conversation
3. failed_axis: Does Assistant 1 fail on the Testing Axis?
   ├──── No  → Provide justification, complete remaining quickly, submit
   └──── Yes → passes_conversation_quality_checks?
               ├──── No  → Provide rejection reason, complete remaining quickly, submit
               └──── Yes → Continue to full evaluation:
                     a. Rate each rubric per model (meets/does not meet)
                     b. Edit rubrics if needed
                     c. Rate rubric quality (5 dimensions)
                     d. Select Best Response
                     e. Rate Overall Satisfaction (x3 assistants)
```

Rubric Quality Ratings Quick Reference

Dimension	Low	Medium	High
Importance	Optional	Important	Essential
Specificity	Not Specific	Somewhat Specific	Very Specific
Atomicity	Not Atomic	Partially Atomic	Fully Atomic
Verifiability	Not Verifiable	Partially Verifiable	Concrete & Checkable
Difficulty	Too Easy	Appropriately Challenging	Too Hard

Quality Checklist
Before submitting:

Checked Testing Axis alignment
Checked conversation quality criteria
Evaluated ALL rubric criteria for ALL 3 model responses
Edited rubrics where needed
Rated rubric quality on all 5 dimensions
Selected Best Response
Rated Overall Satisfaction for all 3 assistants
Added justifications where required