Agent Completion Process Supervision – Part 3 / Antechamber Delivery

Project ID: 683e729ecaf6d16374e6fc5b

Customer Goals and Objectives:

In this project, you will be helping to train an AI chatbot to use different Tools (APIs) to answer the given prompt. These are related to App use

Important Links

- Instructions
- List of tool domains **(07/17)**
- Complete Tool Set Schema **(07/17)**

✅ Project Overview

This project evaluates agent-completion conversations where models use structured tool APIs to fulfill user requests. The focus is on verifying the model's correctness, proper use of tools, and whether the final output aligns with what the user wanted.

Auditors **must only fail for correctness issues**. Domain, objective, and difficulty categorization may be imperfect, but they **must not lead to failure** unless so drastically off that the task cannot be assessed.

🧑‍💻 Contributor Workflow

cb instructions

Step-by-Step Workflow

1. **Create a Natural Scenario**
    - Start by imagining a realistic, complex user situation that would require the help of a virtual assistant.
    - Avoid trivial or artificial setups; scenarios should feel like real conversations someone might have.
2. **Write the System Prompt**
    - Define the assistant's **persona, behavior, tone, context, preferences**, and **rules for tool use**.
    - This prompt determines how the assistant should act. It is unique per task and must follow the structured categories provided in the guidelines.
3. **Write the Initial User Prompt**
    - Initiate the conversation naturally and vaguely, so that the model needs more information/clarification before calling tools.
    - Avoid directly referencing tool names or availability; let the model decide when and how to use tools.
4. **Simulate the Full Conversation**
    - Continue the interaction for **at least 10 user turns**, mixing prompts, tool calls, clarifications, and follow-ups. Include realistic flow, such as:
        - Clarification questions by the agent.
        - Tool errors and recovery.
        - Lazy user behavior or task switching (if relevant).
    - Guide the assistant toward successful task completion or a graceful fallback when the request is infeasible.
5. **Fix Model Mistakes in Real Time**
   At each assistant turn, if the model makes an error (e.g., wrong tool call, missing parameters, hallucinated output), **correct it immediately**.
    - You must:
        - Add error tags (from the allowed list).
        - Provide **critique comments** explaining what went wrong.
        - Enter the **ideal assistant response** (either tool call or text).
    - An exception for this will be the tasks with the category Error Recovery, where the tool call will not be fixed in the first turn - the errors will be flagged, but the model will fail in the first attempt (see error recovery definition) **(06/30)**
6. **Continue Until Task Completion**

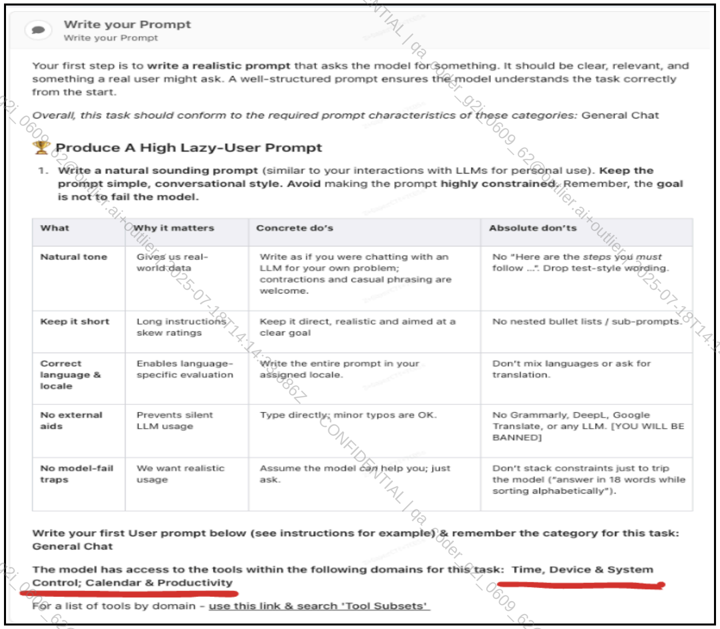
    - Ensure the conversation concludes logically, with the user's request being fulfilled or a clear failure message when appropriate.
    - Strive for conversations that teach the model recovery, adaptability, and multi-step reasoning.

Understanding System Behavior: (06/26)

1. WiFi, Location cannot be turned on when “Low Battery Mode” is turned on
    a. This does not apply to cellular (might change, but we don't think it does)
    b. The pre-seeded system settings might have all of them turned on, this is okay. This rule just applies when the model is trying to change system settings.
        i. The model doesn't have access to the system settings by default. It should use get_system_settings to read them

Audit Workflow

1. Review the pre-provided system settings **(06/24)**
    a. This provides information about the initial state of the user's device. The model can access this information using get_system_settings and modify it using set_system_settings.
    b. If available, the model may or may not use more specific tools (e.g.: get_wifi_status and set_wifi_status to manipulate only wifi settings). This is also correct. **(06/30)**
2. Review what tool subsets are activated in the task. Some tasks have all tools activated, some only have Time, Device & System Control; Calendar & Productivity tools. All tasks have get_system_settings, set_system_settings, get_current_location, & get_current_iso_8601_datetime_with_utc_offset enabled. To understand which tools are enabled for each task - see here in task: **(07/17)**

a. 

Your first step is to **write a realistic prompt** that asks the model for something. It should be clear, relevant, and something a real user might ask. A well-structured prompt ensures the model understands the task correctly from the start.

Overall, this task should conform to the required prompt characteristics of these categories: General Chat

🏆 Produce A High Lazy-User Prompt

1. **Write a natural sounding prompt** (similar to your interactions with LLMs for personal use). **Keep the prompt simple, conversational style. Avoid** making the prompt **highly constrained.** Remember, the **goal is not to fail the model.**

What	Why it matters	Concrete do's	Absolute don'ts
Natural tone	Gives us real-world data	Write as if you were chatting with an LLM for your own problem; contractions and casual phrasing are welcome.	No "Here are the *steps* you *must* follow …". Drop test-style wording.
Keep it short	Long instructions skew ratings	Keep it direct, realistic and aimed at a clear goal	No nested bullet lists / sub-prompts.
Correct language & locale	Enables language-specific evaluation	Write the entire prompt in your assigned locale.	Don't mix languages or ask for translation.
No external aids	Prevents silent LLM usage	Type directly; minor typos are OK.	No Grammarly, DeepL, Google Translate, or any LLM. [YOU WILL BE BANNED]
No model-fail traps	We want realistic usage	Assume the model can help you; just ask.	Don't stack constraints just to trip the model ("answer in 18 words while sorting alphabetically").

Write your first User prompt below (see instructions for example) & remember the category for this task: General Chat

The model has access to the tools within the following domains for this task: Time, Device & System Control; Calendar & Productivity

For a list of tools by domain - use this link & search 'Tool Subsets'

3. Evaluate the System Prompt
4. Evaluate the Prompt
   a. It should sound natural, think of these as spoken prompts
   b. Prompt does not require punctuation
   c. Items can be in quotes "" or not in quotes
   d. Pleasantries like "please" are okay since these are spoken prompts
   e. The prompt should be solvable using the tool set provided
5. Review the model response and evaluate the CB's selection for "Response Type"
6. Evaluate the error type selections and the critique (if applicable)
   a. The contributor must select all applicable error categories
   b. Critic Response: The contributor must mention the correct tool and the tool used. This must be factual
   c. No critic response or reasoning required for error type no_issues

7. Evaluate the contributor’s fixes to the model response
    a. Tool call response:
        i. Tool Selection:
            1. The correct tool should be selected to solve the prompt if the model selects the wrong one
            2. **The tool must clearly be the best tool, you should not be able to argue that the model responded correctly.**
            3. The model should only be using tools within the domains that are assigned at the top of the task UI, see Tool Subsets **(06/26)**
        ii. Parameters
            1. The contributor should correct the inputs for the tools if it is required
            2. If the param type is an array, it must be formatted as an array even if there is only one item
    b. Textual response:
        i. The model text response should summarize the results of the tool calls, answering the prompt or ask for clarification as needed
8. Evaluate the model reasoning
    a. The contributor must describe the thought process behind the steps taken
9. Evaluate if task requirements are followed
10. **IMPORTANT:** Tally Up final score per the Grading Instructions **(07/17)**
    a. Assign rating, error categories and feedback to each taxonomy turn with a user prompt.
    b. Assign the task level rating, any error categories that are not specific to a particular conversation turn and feedback to the task level fields
    c. The task level feedback should be the concatenation of all your conversation turn level feedback. Make sure that turn numbers are clearly mentioned.
    d. Note: Always use internal fields to input your feedback!

Grading Instructions (07/17)

1. This project has turn-based grading enabled now, so each turn in the task can be given its own score
2. We define two types of turns: conversational turns and taxonomy turns
    - Taxonomy turns are the turns you see in the task UI
    - Each conversational turn can be made up of multiple taxonomy turns. A conversational turn begins with a user prompt and ends with the final model response to that user prompt. There may be multiple tool call steps in between.
3. QC should grade each *conversation turn* in the task and also assign a task level score.
    - Add your rating and feedback to the each taxonomy turn with a user prompt
    - The system allows you to add ratings to other taxonomy turns, but DO NOT do this. Stick to turns with user prompts because we only want to assign one rating per conversational turn
4. The rubrics define different grading criteria for rating conversational turns and the task as a whole
5. Instructions for grading conversation turns
    - **General Grading**
        - Grade to the lowest dimension across all rubrics (e.g. if instruction following is a 2, conversation turn task should be rated a 2)
        - If the conversation turn meets any criteria under 1-2 Fail, the conversation turn is a fail.
        - If the conversation turn does not fail and it meets criteria for a 3 on any dimension, then the entire conversation turn is a three.
        - All dimensions must be a 4-5 for the conversation turn to receive a 4-5.
    - **Choosing 1 vs 2 or 4 vs 5**
        - When deciding between a 1 or 2, select a 1 if the attempter put little to no effort
        - When deciding between a 4 or 5, select a 5 if the conversation turn is perfect or impressive
6. Instructions for assigning the task level score are defined in the rubrics.
    - Note: “Task fails if more than 10% of conversational turns fail” should be read as “Task fails if more than 10% of conversational turns fail within the dimension”. Conversation turn fails across different rubric dimensions can’t cause a task level failure.

Rubrics

Task Requirements

Category	Notes	TASK LEVEL GRADING		
		1-2 (Fail)	3 (Okay)	4-5 (Good/Perfect)
Tool Set Usage **UPDATED 06/26**	Ignore [1-40 Tools][40-All Tools] Categories **(06/26)**	**[Fail - Task Specifications Tools Mismatch]** - The task has no tool calls from the assigned domains Note: This is not applicable when the category is "General Chat" or "Infeasible Tool Use" - The task uses tool calls from outside the assigned domains **(06/26)**	N/A	- The task uses at least one tool call from the assigned tool set and none from the other tool sets

Number of User turns	Only user prompt turns are counted as User turns. Empty turns (the result of the interaction of the user when rating tool calls/responses) do not count towards the 10 required turns.	- **[Fail - Not Enough User Turns]** The task has fewer than 10 user turns.	N/A	- The task has at least 10 user turns
Category **UPDATED 06/24**	Here is the list of all the categories.	- **[Fail - Task Category Mismatch]** CB did not follow the task category assigned **Note:** It is fine if the task overlaps with other categories, but the task must align with the assigned categories **(06/24)**	N/A	- CB followed the assigned task category

System Prompt Rubric

Category	Notes	TASK LEVEL GRADING		
		1-2 (Fail)	3 (Okay)	4-5 (Good/Perfect)

System Prompt thoroughness	Building blocks: 1) Context Info 2) Tool-Use rules 3) User Preferences 4) Background Info 5) Tonal Control Read more here	- **[Fail - Missing System Prompt Building Blocks]** Includes < 3 of 5 building blocks		Includes >= 3 of 5 building blocks
Alignment with System Settings **UPDATED 06/26**	**(06/26)** The "System Settings" on top of the system prompt field provides information about the initial state of the user's device - excluding location (sometimes users can use VPN or be in a different location than their phones) The system prompt should not provide any information that contradicts the system state. For example, if the system settings has wifi set to ON, the system prompt must not declare that wifi is OFF.	- **[Fail - System Prompt Contradicts System Settings]** The system prompt contradicts the system settings (excluding location) **(06/26)**	N/A	The information in system prompt aligns with the system settings

Prompt Rubric

Category	Notes	TASK LEVEL GRADING INSTRUCTIONS	CONVERSATIONAL TURN LEVEL GRADING		
			1-2 (Fail)	3 (Okay)	4-5 (Good/Perfect)
Multiturn: Conversational Flow	This is applicable to subsequent prompts in a multiturn conversation. Note the Exceptions and be aware Exception 1, “Lazy User” is a current(6/18) requirement for all tasks. **Exception 1:** when the task asks for a “lazy user behavior”, then the user can provide one or more vague prompts, always within the context of the task. E.g. the model asks for the date and start/end time for a calendar event, and the user only replies with the date, making the model ask again for the time. **Exception 2:** when the task asks for “task switching”, the user can change the subject completely from one prompt to the following one. This new prompt will be outside the conversation general context, but is fine in this type of task.	- Task fails if *more than* 10% of **<u>conversational turns</u>** fail (example: 1/9 conversational turns) - Score of 3 if *less than or equal to* 10% of **<u>conversational turns</u>** fail (example: 1/11 conversational turns) - 4 or 5 if no conversational turns fail	- **[Fail - Repetitive Subsequent Prompt]** Prompt requests for something that is already answered in the previous turns	N/A	- All follow-up prompts are grounded in the conversation and do not digress the conversation into something unrealistic

	Lazy User guidelines				
Lazy User guideline compliance **UPDATED 07/17**	Lazy User guidelines	- Task fails if *more than* 1 **conversational turns** fail - Score of 3 if *exactly* 1 **conversational turn** fails - 4 or 5 if no conversational turns fail	- **[Fail - Lazy User not followed]** The user prompt provides more than one piece of information **(07/17)** **Note:** Not applicable to the first prompt **(06/26)** OR The first prompt does not necessitate the model to ask clarifying questions / search for key information using tool calls	N/A	<= 1 user prompt provides more than one piece of information
Mentions Tool Name	Prompts can coincidentally mention the tool name if it sounds natural. For example, "Can you use Google search to find xyz" is fine, but "Can you use google_search to find xyz" is not	- Task fails if at least one conversational turn fails	- **[Fail - Prompt Mentions Tool Name]** The prompt **mentions** the name of the tool to use. Ex: "Can you use google_search to ..."	N/A	The prompt **doesn't mention** the tool name.
Contrived / Unnatural Prompts **UPDATED 07/17**	**IMPORTANT: Read this to understand what "contrived" and "unnatural" mean** **Prompts should follow Lazy User Persona**	- Task fails if *more than* 10% of **conversational turns** fail (example: 1/9 conversational turns) - Score of 3 if *less than or equal to* 10% of **conversational turns** fail (example: 1/11 conversational turns)	- **[Fail - Contrived / Unnatural Prompt]** Prompt is clearly unrealistic and fantastical. You don't think it's something an end-user would ask. Use good judgment here. - **[Fail - Contrived / Unnatural Prompt]** Prompt is being overly specific and unnatural, to the point where it is weird. Ex: "What are some pizza places in a square root of pi miles radius from me?"	- **[Non-Fail - Somewhat Contrived / Unnatural Prompt]** Prompt is somewhat contrived or is somewhat unnatural, but you could see people asking it - **[Non-Fail - Prompt Includes Some Personal Context]** Includes some unnecessary personal context.	- Prompt is realistic and something an end user would ask

	- 4 or 5 if no conversational turns fail	- **[Fail - Prompt Includes Too Much Personal Context]** Includes too much unnecessary personal context. - **[Fail - Prompt Tailored to a Tool]** Prompt is very clearly targeting a specific tool. Typically, such prompts are designed around the functionality and parameters of the tool that it's targeting Example: I am currently in 650 Townsend, San Francisco, CA 94103. How can I get to Burma Love on Valencia in san francisco by driving - **[Fail - Contrived Scenario]** The CB creates a scenario that is not internally consistent because the user prompt contradicts the information from the tool outputs and/or the system prompt. **(07/17)** **Example:** ● System prompt: "User's home address is xyz" ● Agent calls the current_location tool which outputs "xyz" ● User claims in a prompt that they are not at their home	Example: "Is the gas station open 24x7? its quite late, and I don't want to go there to find out it is closed. Can you help?"	

Model Response and Reasoning Rubric

Category	Notes	TASK LEVEL GRADING INSTRUCTIONS	CONVERSATIONAL TURN LEVEL GRADING		
			1-2 (Fail)	3 (Okay)	4-5 (Good/Perfect)
Response Type Selection	We have 3 response types: 1. text_response 2. tool_call 3. tool_response text_response and tool_call are generated by the model while tool_response is the tool call output	- Task fails if *more than* 10% of **conversational turns** fail (example: 1/9 conversational turns) - Score of 3 if *less than or equal to* 10% of **conversational turns** fail (example: 1/11 conversational turns) - 4 or 5 if no conversational turns fail	- **[Fail - Rewrite Wrong Response Type]** CB may or may not have selected the wrong response type but the following response editing section uses the wrong response type for the rewrite Ex: CB may have recorded a text_response rewrite when it should be a tool_response	- **[Non-Fail - Response Type Selection Wrong]** CB selected the wrong response type but the following response editing section is still the right response type Ex: CB may have incorrectly chosen text_response when it should be tool_response, but the response edit field has a tool_response	- You agree with CB's response type selection
Accuracy **(Applicable to the final model response)** **UPDATED 07/17**	"Trajectory" as used here refers to the agent's tool call chain within a single user-assistant conversation turn, i.e., user prompt → [tool call + tool response] (potentially repeated) → model response **Note:** This does not apply to tool responses. The tool responses should be taken as the SSOT. **(06/26)**	- Task fails if at least one conversational turn fails	- **[Fail - Response Factual Errors]** The model response does not summarize the information from the trajectory accurately or contains unfactual statements - **[Fail - Response Hallucination]** The model response hallucinates information that is	N/A	The model response makes no inaccurate statements

	Major issue: Producing hallucinated information that is factually incorrect or inaccurately summarizing json tool response **(07/17)** **No issue:** Producing information from the model's context/knowledge that is factually correct. Read misstep #3 for more **(07/17)**		factually incorrect (makes up info that can't be grounded in the current turn trajectory or previous turns) **(07/17)**		
Instruction Following / Response Fulfillment **(Applicable to the final model response)** **UPDATED 07/17**	*Rule of thumb for "large" word count requirements: +/-10% is acceptable *Subjective miss* - The prompt asks to explain a photo of a historical event. The response does describe the historical event well but does not mention the year of the event. It objectively answers the question but subjectively one would expect to get the year of the event in the response The model should obey both the system prompt and the user prompt. If the user prompt contradicts the system prompt, the latest context from the user should be prioritized by model. **(06/26)** **Note:** Tonality suggestions are not required to be followed across 100% of prompts unless explicitly requested. For example: **(06/26)** ● System prompt: "The user likes oceans, so employ	- Task fails if *more than* 10% of **conversational turns** fail (example: 1/9 conversational turns) - Score of 3 if *less than or equal to* 10% of **conversational turns** fail (example: 1/11 conversational turns) - 4 or 5 if no conversational turns fail	- **[Fail - Rewrite Explicit Instruction Miss]** 1 or more explicit instructions are not followed - **[Fail - System Prompt Ignored]** The model response does not follow the instructions in the system prompt **(07/17)** - **[Fail - User Prompt Ignored]** When there is a conflict between system prompt and user prompt, the model prefers the system prompt over the user prompt **(07/17)**	- All explicit instructions are clearly followed - **[Non-Fail - Rewrite Subjective Instruction Miss]** Subjectively misses some aspects of fully answering the question	- All explicit instructions are clearly followed - The tool calls fully answer the question

	ocean themed puns” ○ Not all the model responses need to use puns ● System prompt: “The user likes oceans, so make sure to add an ocean themed pun to all of your messages” ○ All model responses should have puns				
Tool Selection **ADDED 06/20**	“Trajectory” as used here refers to the agent’s tool call chain within a single user-assistant conversation turn, i.e., user prompt ⟶ [tool call + tool response] (potentially repeated) ⟶ model response Note that [Incorrect Tool Call] is not applicable if the task category is “Error Recovery” as long as the agent corrects the mistake and recovers from the issue. The model is allowed to self-correct Note: The model should only be using tools within the domains assigned at the top of the task UI, see Tool Subsets **(06/26)**	- Task fails if *more than* 10% of **conversational turns** fail (example: 1/9 conversational turns) - Score of 3 if *less than or equal to* 10% of **conversational turns** fail (example: 1/11 conversational turns) - 4 or 5 if no conversational turns fail	- **[Fail - Incorrect Tool Call]** The tool selected is not suitable for fulfilling the prompt’s request - **[Fail - Irrelevant Tool Call]** At least one inexplicable / irrelevant tool call was made which does not contribute in any way to the trajectory or answering the prompt. **Note:** If the tool call could provide relevant contextual information, but is technically not needed, you can flag that as [Inefficient Trajectory] - **[Fail - Missing Tool Call]** The trajectory does not use a tool call when one should’ve been made **Example:** The agent needed to know what day 2025-07-01 is, but it can’t be deduced from the information available at that point in the trajectory and the agent did not use iso_8601_datetime_with_utc_offset_to_iso_weekday tool	N/A	All tool calls are accurate

			Note: The model does have access to common knowledge and can rely on this knowledge for information that is static in-nature / does not change (e.g. that massachusetts is not close to Michigan, Harvard University is in Boston) **(06/26)**		
Parameter Values **ADDED 07/17**	Note that [Incorrect Parameters] is not applicable if the task category is "Error Recovery" as long as the agent corrects the mistake and recovers from the issue. The model is allowed to self-correct The model can assume some parameters such as limit (how many results to show), radius, calendar event title, etc. It is unusual for the user to mention these parameters, as long as the parameters used by the model are reasonable, this is okay! **(07/17)** Hallucinated or incorrect parameters that are not critical to the end result of the tool calls accuracy / conformity with the users request are not treated as failing issues **(07/17)**	- Task fails if at least one conversational turn fails	- **[Fail - Incorrect Parameters]** The parameter values are incorrect, either because they do not make sense for what the prompt asks or, in case of subsequent tool calls, they don't align with the information from the tool outputs - **[Fail - Hallucinated Parameters]** The parameter values were hallucinated. Meaning, at least one parameter value is not grounded in the information that the agent has access to by the time the tool call was made **Note:** This is applicable even if the parameter values are accurate and typically arises because of [Missing Tool Call]	- **[Non-Fail - Suboptimal Parameters]** The parameter values used were not incorrect, but they could've been better. This is applicable to parameters like "query" in "Google maps" or filter ranges that are too broad, where, even if you did not use the most optimal value, the tool might return the relevant information which the model will then have to intelligently extract, but you could get better results with the optimal parameter **(07/15)**	All parameter values are accurate
Optimal Trajectory **ADDED 06/20**	"Trajectory" as used here refers to the agent's tool call chain within a single user-assistant conversation turn, i.e., user prompt ⟶ [tool call + tool response] (potentially repeated) ⟶ model response	N/A	N/A	- **[Non-Fail - Inefficient Trajectory]** The trajectory is not efficient. It has unnecessary steps that are not really required for answering the prompt but	Trajectory is efficient

				they are not irrelevant.	
Completeness **ADDED 06/20**	"Trajectory" as used here refers to the agent's tool call chain within a single user-assistant conversation turn, i.e., user prompt ⟶ [tool call + tool response] (potentially repeated) ⟶ model response	- Task fails if at least one conversational turn fails	- **[Fail - Incomplete Trajectory]** The trajectory is incomplete and does not provide all the information necessary to answer the prompt	N/A	Trajectory is complete
[Rewrite/SxS] Clearly Worse Than Model Response	*Important Note* - Do not penalize an attempter/tasker for making minimal or no changes to the original response.	- Task fails if at least one conversational turn fails	- **[Fail - Rewrite Worse than Model Response]** The rewrite is objectively worse compared to the original response	- **[Non-Fail - Rewrite Same Quality as Model Response]** The updated response would likely perform about the same overall across the rubric dimensions	- The updated response would clearly perform better overall across the rubric dimensions
Chained vs Parallel tool calls	Read the FAQ section to understand the difference	- Task fails if at least one conversational turn fails	- **[Fail - Dependent Parallel Tool Calls]** Tool calls were made parallelly when they should be chained instead (i.e., when there are dependencies between the tool calls)	- **[Non-Fail - Parallel Tool Calls Opportunity Missed]** Tool calls that could've been called parallelly were split across multiple turns	- Parallel tool calls were made in the same turn - Chained tool calls were appropriately split up across turns
Reasoning Quality **UPDATED 07/17**	**Note:** Error Recovery tasks (task category) will critique the model response - but will not fix/refine it as per category requirements (see error recovery definition) **(06/26)**	- Task fails if *more than* 10% of **conversational turns** fail (example: 1/9 conversational turns) - Score of 3 if *less than or equal to* 10% of **conversational turns** fail (example: 1/11 conversational turns)	- **[Fail - Reasoning Missing]** The model response was marked as incorrect, but reasoning is missing - **[Fail - Reasoning Uses Past Tense]** Reasoning uses past tense to describe the action it's going to take ("I used calendar tool" rather than "I will use calendar tool") **Note:** Chronologically, the reasoning comes before the tool call or text response from the	N/A	- Reasoning is clear, specific and appropriately explains the action it's going to take

		- 4 or 5 if no conversational turns fail	same turn. Actions from the previous turns can be described using past tense **(07/10)** - **[Fail - Incoherent Reasoning]** Reasoning is not coherent with the tool calls - **[Fail - Reasoning Poor Framing]** The reasoning is not written as the model's internal thoughts in first person perspective before making the tool calls / responding to the user. For example: "The model should do xyz" rather than "I will do xyz" **(07/17)**		
Redundant Tool Calls	Exception: the model has a dataset for each task, where some settings are set by default in specific states (e.g. Wifi turned off, low-mode battery turned on, etc.). The model will try to adjust settings to the right state to perform the user's expected actions, like turning Wifi on before searching for a place. Sometimes more than one tool call needs to be made until settings are correctly set. Do not penalize for this setting adjustment.	- Task fails if *more than* 10% of **conversational turns** fail (example: 1/9 conversational turns) - Score of 3 if *less than or equal to* 10% of **conversational turns** fail (example: 1/11 conversational turns) - 4 or 5 if no conversational turns fail	- **[Fail - Redundant Tool Calls]** The model makes a tool call to retrieve info that is already known from the previous turns	N/A	- There are no redundant tool calls

Error Type Selection and Critique

Category	Notes	TASK LEVEL GRADING INSTRUCTIONS	CONVERSATIONAL TURN LEVEL GRADING		
			1-2 (Fail)	3 (Okay)	4-5 (Good/Perfect)
Error Type Selection	**Note:** Error Recovery tasks (task category) will critique the model response - but will not fix/refine it as per category requirements (see error recovery definition) **(06/26)**	- Task fails if *more than* 10% of **conversational turns** fail (example: 1/9 conversational turns) - Score of 3 if *less than or equal to* 10% of **conversational turns** fail (example: 1/11 conversational turns) - 4 or 5 if no conversational turns fail	- **[Fail - Incorrect Error Type]** Wrong error type is selected - **[Fail - Error Type Not Selected]** CB missed selecting at least one error type	N/A	- You agree that the error types chosen are accurate and cover everything wrong with the model responses
Critique		- Task fails if *more than* 10% of **conversational turns** fail (example: 1/9 conversational turns) - Score of 3 if *less than or equal to* 10% of **conversational turns** fail (example: 1/11 conversational	- **[Fail - Missing Critique]** Critique is missing for an error type that's selected - **[Fail - Incorrect Critique]** Critique is inaccurate	N/A	- Critique appropriately justifies what was wrong with the model response, explaining the reasoning behind each of the error types selected

		turns) - 4 or 5 if no conversational turns fail			

Core Dimensions (Applicable across all CB generated / edited content)

Category	Notes	TASK LEVEL GRADING INSTRUCTIONS	CONVERSATIONAL TURN LEVEL GRADING		
			1-2 (Fail)	3 (Okay)	4-5 (Good/Perfect)
Original Work *(QC Only)*	**On AI Usage:** You should assume using LLMs to generate portions of the contributor generated content is unsanctioned unless otherwise indicated in the project instructions or the spec doc. When it is obvious that the CB used LLMs, fail the task with a rating of 1. If you have a suspicion that they used LLMs, but there is plausible deniability – meaning, you could reasonably argue that someone who was not using LLMs could have also done the same thing – use the error category **[Non-Fail - Suspected LLM Cheating]** but this should NOT affect the task rating. **On Plagiarism:** ● For prompts: If you come across a coding prompt that asks for something unoriginal, like for example "Can you create tic-tac-toe in Pygame", this is not necessarily flagged under plagiarism. This prompt is generic enough that you will see it in multiple places online, but unless project instructions say otherwise, the idea here is to catch fairly unique prompts that	- Task fails if at least one conversational turn fails	- **(1) [Fail - LLM Cheating]** The task content contains undeniable evidence of unsanctioned LLM usage by the CB - **(1) [Fail - Plagiarism]** Task contains plagiarized content	N/A	- You don't think the CB engaged in unsanctioned LLM usage - Task contains no plagiarism - **[Non-Fail - Suspected LLM Cheating]** You have a suspicion that the CB engaged in LLM usage for coming up with parts of the task, but there's plausible deniability

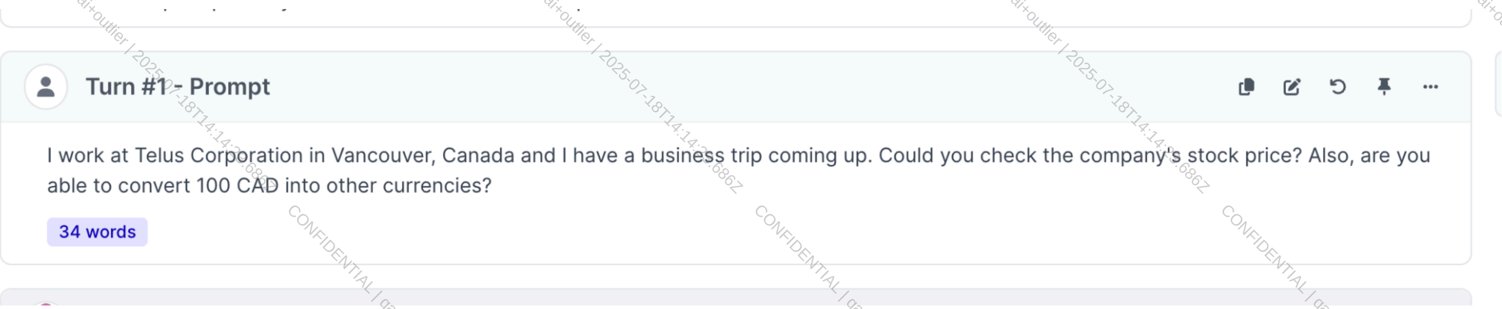
	were blatantly copied (or just rephrased without meaningfully modifying the problem) from a specific source on the internet / textbook. • For everything else: Too much text is copied verbatim from a different source without citation.				
Harmful Content	See project specific or general guidelines on what constitutes "harmful content" Beyond project specific, Scale's safety team enumerates two types of harmful content: 1. Content harms - unsafe text (bigotry, conspiracy theories) 2. Facilitations harms - text that enables unsafe behavior (how to make a bomb) Be on the lookout for both!	- Task fails if at least one conversational turn fails	- **(1) [Fail - Harmful Content]** The content contains any harmful content	N/A	- Content does not contain or asks about harmful content
Native Fluency	Things to consider include writing style, tone, word choice, verbosity, and awkward sentence structure. Examples include: word for word translation from another language "Incorrect word order: In German, saying ""Ich habe gekauft das Buch"" instead of ""Ich habe das Buch gekauft"" (I bought the book). Overuse of pronouns: In Spanish, saying ""Yo voy a la tienda, y yo compro	- Task fails if at least one conversational turn fails	- **[Fail - Lacks Native Fluency]** Writing is not that of a native speaker in the specified language and locale.	- **[Non-Fail - Some Fluency Errors]** Writing is mostly that of a native speaker, but there is 1 strange phrase that gives you pause.	- Native-level writing for specified language and locale.

	pan"" instead of just ""Voy a la tienda y compro pan"" (I go to the store and buy bread). Literal translations of idioms: In French, saying ""Il pleut des chats et des chiens"" (It's raining cats and dogs) instead of the correct idiom ""Il pleut des cordes"" (It's raining ropes). Formal/informal confusion: In Japanese, using casual language (友達語) when a more formal style (敬語) is appropriate, or vice versa."				
Spelling / Grammar / Formatting	NOTE: Scale these standards to the length of content as appropriate. I.E. if the content is very short (a paragraph or less) you may grade more harshly "Grammar" as used here constitutes punctuation, syntax, wording, sentence, word choices, etc An egregious error is something that changes the meaning of what's written or is a completely jumbled spelling or sentence	- Task fails if at least one conversational turn fails	- **[Fail - Many Egregious Spelling Errors]** Has 4 or more egregious spelling errors - **[Fail - Broken Formatting]** The response contains broken formatting such as broken formatting for a list or broken markdown Note: For spelling, grammar, and punctuation, errors accumulate across the entire task. For example, if the prompt had	- **[Non-Fail - Minor Grammar and Punctuation Errors]** Has 4 or more spelling (minor), grammar, and punctuation errors - **[Non-Fail - Some Egregious Spelling Errors]** Has at most 3 egregious spelling errors - **[Non-Fail - Subpar Formatting]** Minor formatting issues such as multiple new lines between content	- **(5)** There are no easily discernible errors - **[Non-Fail - Minor Grammar and Punctuation Errors] (4)** Up to 3 minor errors in spelling / grammar / formatting

			2 egregious spelling errors and the response had 2 egregious spelling errors, we'd fail the task		
Clarity / Structure	Consider whether the content can be improved by altering word choice/syntax, sentence structure, or idea organization. Note: For Reasoning tasks tasks this would require outlining the logical steps required to reach a given conclusion	- Task fails if at least one conversational turn fails	- **[Fail - Major Clarity Issues]** Content is extremely difficult to follow or is unclear.	- **[Non-Fail - Minor Clarity Issues]** Content makes sense but has some minor clarity issues.	- Content is clear and makes sense
Repetitiveness / Relevance	NOTE: Scale these standards to the length of content as appropriate. I.E. if the content is very short (a paragraph or less) you may grade more harshly *Rule of thumb* - You could delete the irrelevant or repetitive material and it would not materially detract from the content	- Task fails if at least one conversational turn fails	- **[Fail - Repetitive Content]** The content contains unnecessary repetition, having 3 or more sentences that express the exact same idea - **[Fail - Not Relevant (4+)]** The content contains 4 or more irrelevant sentences	- **[Non-Fail - Not Relevant]** The content contains 3 irrelevant sentences	- The content does not contain unnecessary repetition, having 2 or fewer sentences that express the same idea - The content contains 2 or fewer irrelevant sentences

Screenshots: Which fields to grade

1. Prompt:

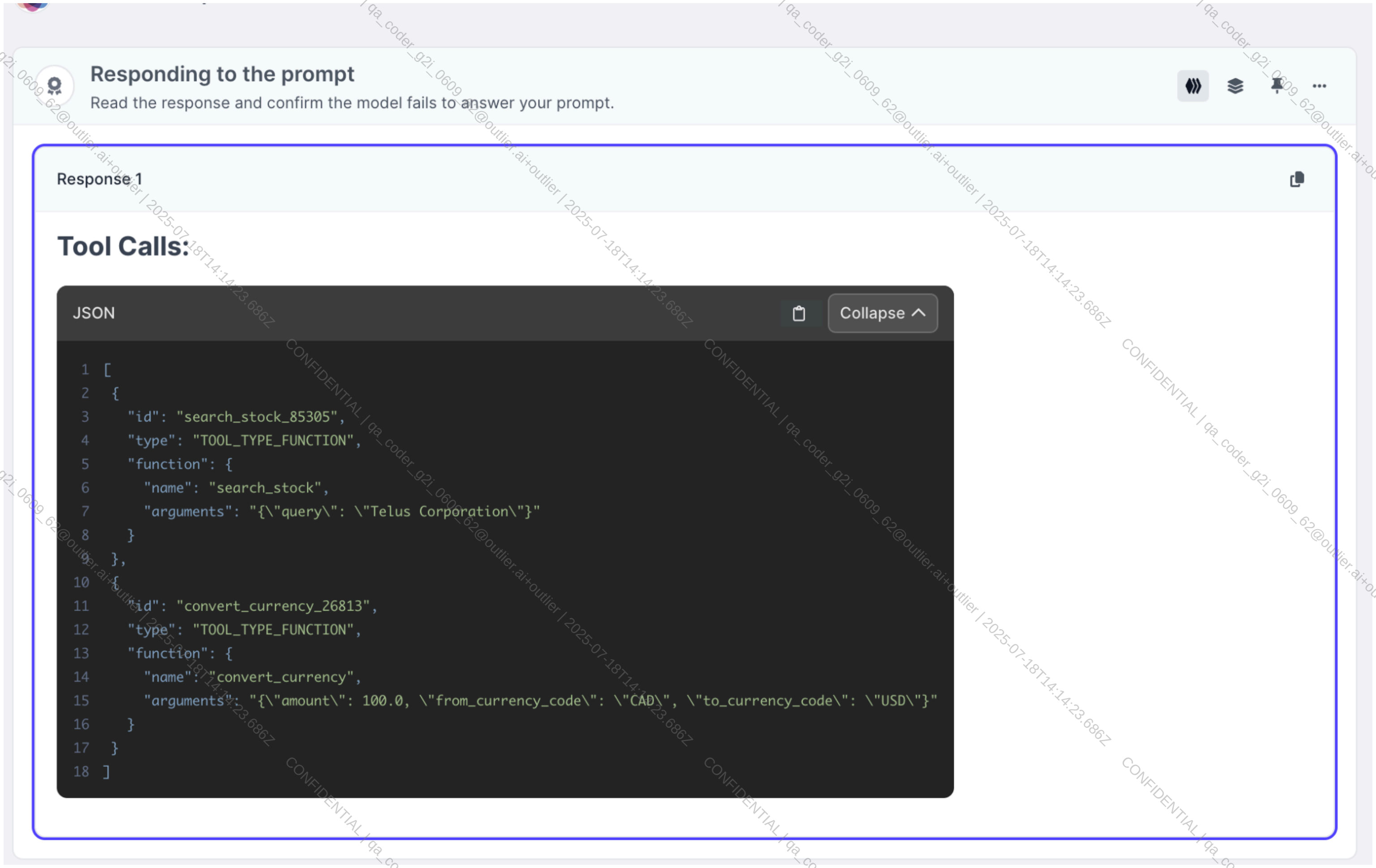


I work at Telus Corporation in Vancouver, Canada and I have a business trip coming up. Could you check the company's stock price? Also, are you able to convert 100 CAD into other currencies?

   a.

2. Response:

a. Original Response: This is the first message in the response to the prompt



Responding to the prompt

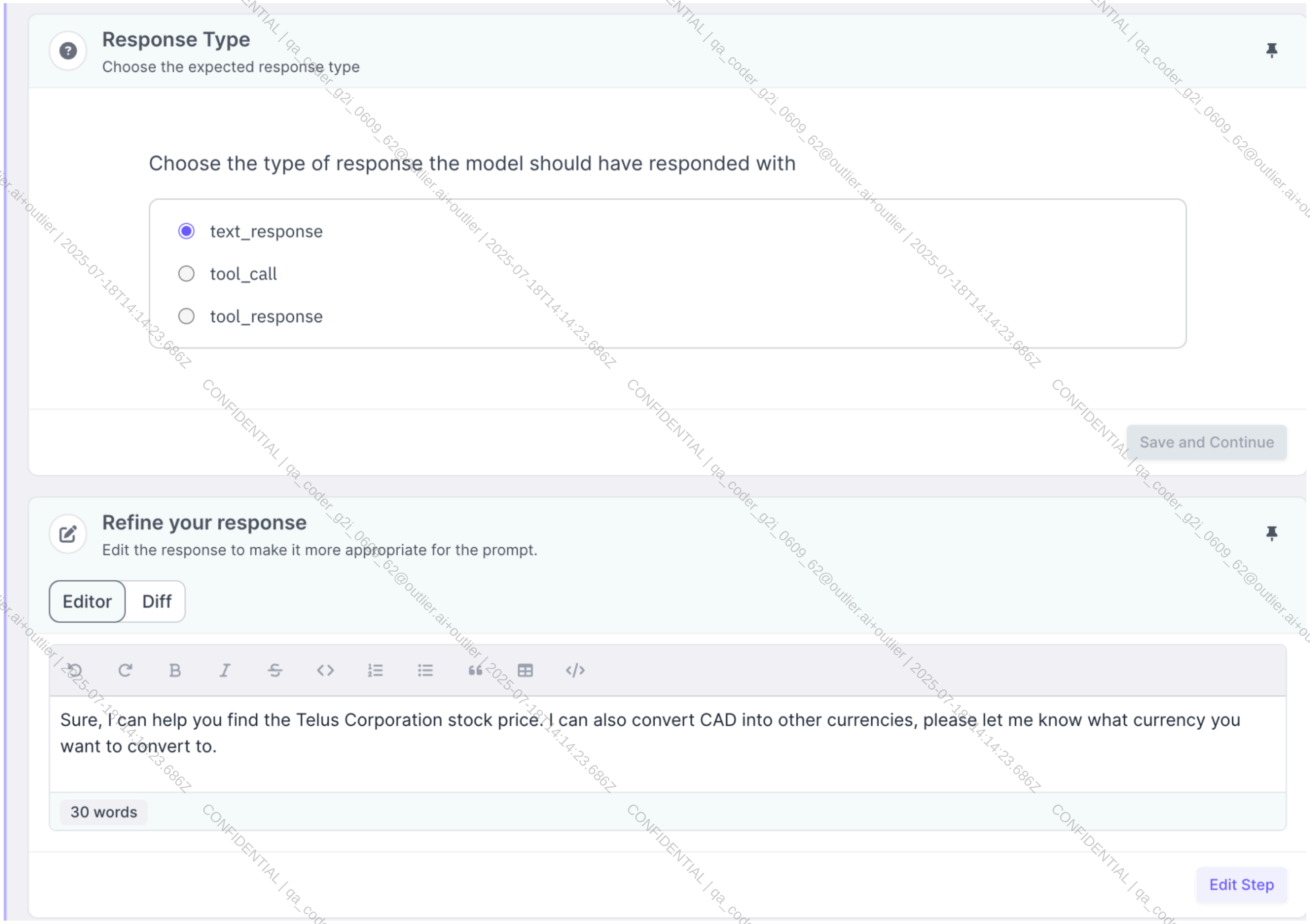
Read the response and confirm the model fails to answer your prompt.

Response 1

Tool Calls:

```
[
 {
   "id": "search_stock_85305",
   "type": "TOOL_TYPE_FUNCTION",
   "function": {
     "name": "search_stock",
     "arguments": "{\"query\": \"Telus Corporation\"}"
   }
 },
 {
   "id": "convert_currency_26813",
   "type": "TOOL_TYPE_FUNCTION",
   "function": {
     "name": "convert_currency",
     "arguments": "{\"amount\": 100.0, \"from_currency_code\": \"CAD\", \"to_currency_code\": \"USD\"}"
   }
 }
]
```

b. Corrected Response: This is the second message after the response



Response Type

Choose the expected response type

Choose the type of response the model should have responded with

● text_response

○ tool_call

○ tool_response

Refine your response

Edit the response to make it more appropriate for the prompt.

Sure, I can help you find the Telus Corporation stock price. I can also convert CAD into other currencies, please let me know what currency you want to convert to.

30 words

i.

3. Response Critique: (**This is the same as the prior App Tool Use Project)**
    a. Error Types: The error type from the mode, these are all in the task.
    b. Critic Comments: Why the model response is bad.

    c. Reasoning: The reasoning behind the corrected response

4. Why are some prompts blank?
    a. Prompts are blank when:
        i. The model has to execute a tool call
        ii. The model has to summarize the execution of a tool call