

# Instructions

In this task, you will be provided with a **Prompt** from a user (e.g., a question, instruction, statement) to an AI chatbot along with two potential machine-generated **Responses** to the Prompt.

Your job is to assess which of the two Responses is better for the Prompt, considering the following for each Response:

<p><b>Helpfulness:</b> To what extent does the Response provide useful information or satisfying content for the Prompt?</p> <p>Responses should:</p> <ul style="list-style-type: none"><li>▪ <b>Address the intent of the user's Prompt</b> such that a user would not feel the Prompt was ignored or misinterpreted by the Response.</li><li>▪ <b>Provide specific, comprehensive, and up-to-date information</b> for the user needs expressed in the Prompt.</li><li>▪ <b>Be sensible and coherent.</b> The response should not contain any nonsensical information or contradict itself across sentences (e.g., refer to two different people with the same name as if they are the same person).</li><li>▪ <b>Adhere to any requirements indicated in the Prompt</b> such as an explicitly specified word length, tone, format, or information that the Response should include.</li><li>▪ <b>Not contain inaccurate, deceptive, or misleading information</b> (based on your current knowledge or quick web search - you do not need to perform a rigorous fact check)</li><li>▪ <b>Not contain harmful, offensive, or overly sexual content</b></li></ul> <p>A Response may sometimes intentionally avoid or decline to address the question/request of the Prompt and may provide a reason for why it is unable to respond. For example, "Sorry, there may not be a helpful answer to this question." These responses can be considered helpful in cases where an appropriate helpful response to the Prompt does not seem possible.</p>	<p>Rating scale:</p> <ul style="list-style-type: none"><li>▪ <b>Not at All Helpful:</b> Response is useless/irrelevant, contains even a single piece of nonsensical/inaccurate/deceptive/misleading information, and/or contains harmful/offensive/overly sexual content.</li><li>▪ <b>Slightly Helpful:</b> Response is somewhat related to the Prompt, does not address important aspects of the Prompt, and/or contains outdated information.</li><li>▪ <b>Somewhat Helpful:</b> Response partially addresses the intent of the Prompt (most users would want more information), contains extra unhelpful information, and/or is lacking helpful details/specifics.</li><li>▪ <b>Very Helpful:</b> Response addresses the intent of the Prompt with a satisfying response. Some users might want a more comprehensive response with additional details or context. It is comparable to a response an average human with basic subject-matter knowledge might provide.</li><li>▪ <b>Extremely Helpful:</b> Response completely addresses the intent of the Prompt and provides helpful details/context. It is comparable to a response a talented/well-informed human with subject-matter expertise might provide.</li></ul>
<p><b>Presentation:</b> To what extent is the content of the Response conveyed well?</p> <p>Responses should:</p> <ul style="list-style-type: none"><li>▪ <b>Be organized in a structure that is easy to consume and understand.</b> Flowing in a logical order and makes good use of formatting such paragraphs, lists, or tables.</li><li>▪ <b>Be clearly written in a polite neutral tone</b> that is engaging, direct, and inclusive. The tone should not be <i>overly</i> friendly, salesy, academic, sassy, or judgmental in a way that most users would consider to be off-putting or overdone.</li><li>▪ <b>Have consistent style with natural phrasing and transitions</b> as if composed by a single talented human.</li><li>▪ <b>Not be rambling, repetitive, or contain clearly off-topic information.</b> Similar information should not be repeated multiple times. It is harder for users to consume the helpful information in a response if there is repetitive or less helpful information mixed into the response.</li><li>▪ <b>Not include notable language issues or grammatical errors</b></li></ul>	<p>Rating scale:</p> <ul style="list-style-type: none"><li>▪ <b>Poor:</b> Response is poorly written or has notable structural, formatting, language, or grammar issues. Or Response has an awkward or inappropriate tone. Or the Response repeats similar information. Or only a small portion of the Response contains helpful information.</li><li>▪ <b>Adequate:</b> Response could have been written/organized better or may have minor language/grammar issues. A minimal amount of less helpful information may be present. Users would still feel the content of the Response was easy to consume.</li><li>▪ <b>Excellent:</b> Response is very well written and organized. Sentences flow in a logical order with smooth transitions and consistent style. The content of the Response is conveyed in a way that is comparable to a response a talented human might produce.</li></ul>

Overall, you should consider both factors in your SxS rating of which response is better. A more concise response presenting the most helpful information directly and clearly is usually better than a longer response that may be harder to consume and/or contains clearly off-topic information. Responses with Poor Presentation (e.g., rambling, inappropriate tone) should play a significant role in your assessment of which side is better. It may help to imagine the user chatting with a real person and consider which Response most users would prefer to receive from a real person.

## Task Description

In this task, you will be shown a **Conversational Web Search**. A Conversational Web Search is a text conversation directed by a user chatting with an AI chatbot that will reply with web search results or a text response to each message the user sends. This is a conversation, so the context of the messages sent earlier in the conversation is relevant.

You will be provided with context about the state of the Conversational Web Search, including the following:

- **User Messages:** A message issued by the user in the Conversational Web Search. The most recent User Message is referred to as the Current User Message.
- **Responses:** Text responses or web search results from the AI chatbot to the User Message. *The Responses may or may not be visible in the task.*
- **Approximate User Location:** The user's approximate location when the Current User Message was issued.

You will also be shown a collection of **Follow-Up Message suggestions**, or messages the user can issue to find new dimensions, alternatives, next steps, related ideas, or new starting points for their topic of interest. Your job will be to **rate the quality of some Follow-Up Message suggestions** provided to the user for what they may want to ask next. The suggestions should be interesting and exploratory, not a rephrase of the Current User Message; instead they should be ideas about other aspects of the topic the user can ask about.

## How to Rate Follow-Up Messages

- 1. Understand the Current User Message and current state of the Conversational Web Search.**
  - Review the Current User Message and additional conversation context to better understand the information needs of the user. You are encouraged to research the messages if the user intent is unclear.
  - Examine the Current User Message and select the appropriate option about whether it is seeking content that is: (i) Appropriate for All, (ii) Racy or Not-for-Everyone, (iii) Porn or Upsetting-Offensive. If the intent of the message is unclear, mark it as Appropriate for All.
- 2. Assign spelling, foreign language, and appropriateness flags:**
  - **Misspelled:** The Follow-Up is misspelled or has grammar issues.
  - **Foreign Language:** The Follow-Up is in a foreign language for the task location or the Current User Message.
  - **Not Appropriate for All:** The Follow Up Message is inappropriate or unsuitable for some environments or some group of users, including children or other sensitive users.
- 3. Identify the Duplicate Follow-Ups — both to the Current User Message or the other Follow-Ups:**
  - Two Follow-Up Messages are *Duplicates* of each other when either of the following is true:
    - A single complete and concise answer can sufficiently answer both messages.
    - A human could take the answer for one of the messages and easily determine the answer for the other without any extra information.
  - To mark a Follow-Up as a *Duplicate* to the Current User Message or Previous Messages, mark Yes to the extra question.
  - To mark a Follow-Up as a *Duplicate* to another Follow-Up, click on **'Select duplicate Follow-Ups.'**
- 4. Evaluate each Follow-Up.**
  - Follow-Up Messages are suggestions intended to help users find new dimensions, alternatives, next steps, related ideas, or new starting points for their topic of interest. Assume users have already issued the Current User Message and have read the Current Response (if available), and now would like to issue a new message.
  - The checkboxes correspond to the following Follow-Up characteristics:
    - **Off-topic:** Is the Follow-Up Message irrelevant to the rest of the Conversational Web Search?
    - **Exploratory:** Does the Follow-Up Message explore an interesting new dimension of the Conversational Web Search topic(s)?
    - **Intriguing:** Would the Follow Up Message inspire most users to want to know more about its topic?
    - **Ambiguous:** Is the meaning of the Follow-Up Message unclear?
  - Rate the Usefulness of the Follow-Up.
    - **Very Useful:** The Follow-Up suggestion is **interesting or useful for most users** who have already issued the Current User Message and Previous Messages (if present).
    - **Useful:** The Follow-Up suggestion is **interesting or useful to many users** who have already issued the Current User Message and Previous Messages (if present).

- **Somewhat Useful:** The Follow-Up suggestion **is interesting or useful for some users** who have already issued the Current User Message and Previous Messages (if present). This could be because the Follow-Up suggestion doesn't offer enough that is sufficiently different from the Current User Message or because the Follow-Up suggestion drifts too far from the topic of interest.
- **Not Useful:** The Follow-Up suggestion **is not interesting or useful for users** who have already issued the Current User Message and Previous Messages (if present). Follow-Ups which are **Too similar** to the Current User Message are Not Useful.

#### 5. Evaluate the sides holistically.

- **Assign an Overall SxS Rating:** A side with more Follow-Ups is not necessarily better than a side with fewer. Follow-Ups which are too similar to the Current User Message or lead to similar sets of results as the Current User Message are not only unhelpful but also a waste of time for users. Please think carefully about whether the Follow-Ups take users in helpful new directions, given that they have already issued the Current User Message and seen its results.
- **Provide feedback about whether the Follow-Ups on each side provide a combination of breadth, depth, and interestingness as a group.** Users will see the Follow-Up Messages together, so it is important that they be useful together, not just individually.
- **Provide feedback about whether the Follow-Ups on each side provide a sufficient variety of directions in which to take the Conversational Web Search if presented as a group.** Diversity and coverage of possible conversational directions are important aspects of a good group of Follow-Up Messages.

## Instructions

In this task, you will be provided with a long Machine-Generated Response to a user's question(s) along with several Target Sentences and their cited Evidence. The Target Sentences should be based on information found within the cited Evidence, however it will often either misrepresent the information or will provide additional information not found in the Evidence.

The purpose of this task is to detect those misrepresentations and/or the additional information as you compare the Target Sentence to the Evidence. For task purposes, consider the Evidence reliable. Your job is to read through the Response and then rate the task sentence-by-sentence:

**A. To what extent can the Target Sentence be verified by its Evidence?**

- Only use the cited Evidence.
- Look for information in the highlighted sentence that is "**Unsupported**" (not contained in the evidence) or "**Contradictory**" (misrepresents the evidence).
- The Target Sentence can be "**Fully**" supported by its Evidence even if
  - the sentence rephrases and/or combines information from several pieces of Evidence;
  - the sentence leaves out some information from its Evidence.

**B. Mark whether there is irrelevant Evidence cited by the Target Sentence.**

- An irrelevant piece of Evidence does not support or contradict the Target Sentence in any way.

Note that the two questions above are purely about a single, highlighted Target Sentence and the Evidence it cites each time. The other sentences in the Response should be used as context to understand the Target Sentence.

### Examples

## Machine-Generated Response:

*Machine:* Yes, Lisinopril is a commonly prescribed medication.

*User:* How long does it take to work?

*Machine:* Lisinopril may take 2 to 4 weeks to work fully. [1]

*User:* Will it help with my hypertension?

*Machine:* Yes, 20mg Lisinopril can lower blood pressure by 10mm Hg! [5,6]

*User:* Again, how soon will it work?

*Machine:* Lisinopril [1] may take 2 to 4 weeks to achieve full benefit [5].

There are several reasons why doctors would prescribe Lisinopril, even if patients do not feel sick, for example, to improve a patient's outcomes after a heart attack. [1]

The average cost of Lisinopril is around \$16 for 30 20mg pills. [2]

There are larger packs available too, which may make the medication cheaper to take.

*User:* Will insurance pay for it?

*Machine:* It's on the list of covered medications at Anthem [1,3] and HealthAmerica.

BlueCross also seems to cover it [3,4]. Almost every American takes it! [7,8]

<p style="text-align: center;"><b>Target Sentence</b> (highlighted) and <b>Evidence</b></p>	<p style="text-align: center;"><b>A. To what extent can the Target Sentence be verified by its Evidence?</b></p>	<p style="text-align: center;"><b>B. Mark whether there is irrelevant Evidence cited by the Target Sentence.</b></p>
<p><i>Machine:</i> Yes, Lisinopril is a commonly prescribed medication.</p> <p><i>User:</i> How long does it take to work?</p> <p><i>Machine:</i>  Lisinopril may take 2 to 4 weeks to work fully. [1] ...</p> <div style="border: 1px solid gray; padding: 10px; margin-top: 10px;"> <p><b>Evidence:</b></p> <p>[1] <i>Lisinopril - Rx Outreach</i>   Most people with high blood pressure do not feel sick. For the treatment of high blood pressure with Lisinopril, it may take 2 to 4 weeks before you get the full benefit of this medication. For the treatment of heart failure, it may take weeks to months before you get the full benefit of this medication.  <a href="http://rxoutreach.org/monograph/lisinopril/">rxoutreach.org/monograph/lisinopril/</a></p> </div>	<p><b>Unsupported:</b> Some important information is not contained in the cited Evidence; no contradiction.</p> <p><i>Explanation:</i> We do not know what Lisinopril is to be used for. The Target Sentence omits the condition for the provided information (2 to 4 weeks), which is "high blood pressure". We cannot assume that this condition is indeed applicable. Thus, even the important information is not sufficiently supported in this case.</p>	<p>None</p> <p><i>Explanation:</i> The single piece of Evidence is relevant.</p>
<p>...<i>User:</i> Will it help with my hypertension?</p> <p><i>Machine:</i> Yes, 20mg Lisinopril can lower blood pressure by 10mm Hg! [5,6] ...</p> <div style="border: 1px solid gray; padding: 10px; margin-top: 10px;"> <p><b>Evidence:</b></p> <p>[5] <i>Lisinopril - Rx Outreach</i>   Studies have shown that Lisinopril at 20mg/daily can lower blood pressure by 10-30mmHg after about 3 weeks of treatment.  <a href="http://rxoutreach.org/monograph/lisinopril/">rxoutreach.org/monograph/lisinopril/</a></p> <p>[6] <i>Facts about Hypertension</i>   <a href="http://cdc.gov">cdc.gov</a>   But millions of adults in the U.S. have hypertension (high blood pressure) – many between their late 30's to early 40's are starting to have this medical condition. Hypertension is a blood pressure measurement with the systolic (top) number being higher than 130 mm Hg or the diastolic (bottom) number being over 80 mm Hg.  <a href="http://www.cdc.gov/bloodpressure/facts.htm#:~:text=Nearly_half_of_adults_in,are_taking_medication_for_hypertension.">www.cdc.gov/bloodpressure/facts.htm#:~:text=Nearly_half_of_adults_in,are_taking_medication_for_hypertension.</a></p> </div>	<p><b>Partially:</b> All important information is supported by the cited Evidence (though some trivial/ minor information is not supported); no contradiction.</p> <p><i>Explanation:</i> A change in wording is acceptable as long as it does not affect the meaning. However, the omission of "daily" and "up to 30 mmHg" makes this sentence only partially supported.</p>	<p>[6] - not relevant</p> <p><i>Explanation:</i> Evidence [6] does not say anything about Lisinopril lowering blood pressure. Thus is it not relevant to the Target Sentence, even if it may give helpful contextual information.</p>

...User: Again, how soon will it work?

Machine: [Lisinopril][1] may take 2 to 4 weeks to achieve full benefit [5].

...

**Evidence:**

[1] *Lisinopril - Rx Outreach* | Most people with high blood pressure do not feel sick. For the treatment of high blood pressure with Lisinopril, it may take 2 to 4 weeks before you get the full benefit of this medication. For the treatment of heart failure, it may take weeks to months before you get the full benefit of this medication.  
[rxoutreach.org/monograph/lisinopril/](http://rxoutreach.org/monograph/lisinopril/)

[5] *Lisinopril – Rx Outreach* | Studies have shown that Lisinopril at 20mg/daily can lower blood pressure by 10-30mmHg after about 3 weeks of treatment.  
[rxoutreach.org/monograph/lisinopril/](http://rxoutreach.org/monograph/lisinopril/)

**Fully:** Everything in the Target Sentence can be verified by the cited Evidence.

*Explanation:* The Target Sentence is the same as seen in the first example, but this time the context is different. Now we know that it is used for the treatment of high blood pressure. For this condition, Evidence [1] lists 2-4 weeks as the time Lisinopril takes to provide full benefit. That piece of evidence alone would be enough.

You should evaluate Evidence [1] like any other. Evaluate the entire sentence against all the cited evidence.

None

*Explanation:* Both [1] and [5] are relevant, as both of them support the Target Sentence. It is OK that [5] does not provide an additional support.

Machine: There are several reasons why doctors would prescribe Lisinopril, even if patients do not feel sick, for example, to improve a patient's outcomes after a heart attack. [1] ...

**Evidence:**

[1] *Lisinopril - Rx Outreach* | Most people with high blood pressure do not feel sick. For the treatment of high blood pressure with Lisinopril, it may take 2 to 4 weeks before you get the full benefit of this medication. For the treatment of heart failure, it may take weeks to months before you get the full benefit of this medication.  
[rxoutreach.org/monograph/lisinopril/](http://rxoutreach.org/monograph/lisinopril/)

**Unsupported:** Some important information is not contained in the cited Evidence; no contradiction.

*Explanation:* Evidence 1 lists different conditions (hypertension, heart failure), but it does not provide information about the treatment of heart attacks. This information was added by the machine.

None

*Explanation:* Evidence 1 still provides a little support here (that there are "several reasons"), so it is relevant and should not be marked here.

*Machine:* The average cost of Lisinopril is around \$16 for 30 pills [2]; check out [GoodRX]([www.goodrx.com/prices](http://www.goodrx.com/prices)) for some prices. [9] ...

**Evidence:**

[2] Without insurance, the median cost of lisinopril ranges from \$15.25 to \$17.10 (20mg, 30 capsules). Lisinopril treats high blood pressure and improves survival after a heart attack. [January 1 2023, 3 months ago]

[9] *Prescription Prices, Coupons and Pharmacy Information* | Stop paying too much for prescriptions Compare prices and save up to 80% [www.goodrx.com/prices](http://www.goodrx.com/prices)

[12] Lisinopril is priced at \$69.50 (20mg, 30 capsules). [October 1 2018, 5 years ago]

**Partially:** All important information is supported by the cited Evidence (though some trivial/ minor information is not supported); no contradiction.

The website link ("/prices") and Evidence [9] both sufficiently support the second part of the Target Sentence, namely that the linked website contains related prices. Do not visit the landing page of any links to attempt gather more information.

*Explanation:* "Median" is incorrectly paraphrased to "average", and the Target Sentence leaves out an important condition (20mg pills). Yet, it provides an approximate and useful answer to the core of the question. The detail of "is around \$16" by itself would be fully supported by the Evidence ("around" indicates an approximation of the cost).

[12] - not relevant

*Explanation:* Evidence [12] does not contradict [2] because it is old and superseded by the newer evidence. [12] is not relevant because it is too old to be reliable.

# Instructions

In this task, you will be provided with a **Prompt** from a user (for example, a question, instruction, statement) to an AI chatbot along with machine-generated **Responses** to the Prompt - which can sometimes include inaccurate pieces of information.

Your job is to assess the **level of accuracy** of the factual information contained in the Responses. Factual information refers to information or statements in responses that could be considered correct or incorrect. Many statements do not contain factual information such as generic statements (for example "hello", "hope you are doing well"), subjective statements, and fictional content.

1. Read the Response carefully while noting the factual information contained in the Response. Identify the most important pieces of factual information in the Response in the context of addressing the Prompt and any suspicious factual information that you think is more likely to be inaccurate (such as any dates/numbers provided by machine-generated responses should be considered suspicious).
2. Select whether this response is actually **attempting to convey any factual information**. Some responses do not - for example, "Once upon a time, a magical lizard went walking through the forest".
3. Use a search engine to research the important and suspicious factual information you identified. Refer to multiple high quality, trustworthy sources.
4. Select the category that best represents the level of accuracy of the response:
  - **Completely accurate:** All the factual information in the Response is verifiably correct or consistent with an established expert consensus (medical, scientific, historical, etc.).
  - **Reasonably accurate:** The most important factual information in the Response is accurate or would widely be viewed as accurate by non-experts. However, the Response may include minor inaccuracies in less important factual information or contain factual information presented in a way that could potentially be misleading.
  - **Questionably accurate:** At least one piece of important factual information is generally considered debatable or controversial.
  - **Not accurate:** At least one piece of important factual information is verifiably incorrect or contradicts an established expert consensus (medical, scientific, historical, etc.).
  - **Can't confidently assess:** The Response is unclear or it is difficult to sufficiently determine the accuracy of at least one piece of important factual information.

## Guidelines for Canned Responses

Some responses are "**canned responses**" - predefined responses that are not actually attempting to answer the question, but signal that the chat service isn't able to. Examples include:

- "I'm a language model and don't have the capacity to help with that."
- "I'm sorry, I don't understand what you're asking me to do. Can you please rephrase your question?"
- "That feature is not available right now."

Note that these are all attempting to avoid the question, not answer it. You will be asked to indicate if these exist in the responses.

## Examples

Prompt	Response	Rating Explanation
Who is Michael Jackson?	Michael Jackson was an American singer, songwriter, and dancer who was the most popular entertainer in the world in the early and mid-1980's. Even today he is widely regarded as the "King of Pop." Born on August 29, 1958 in Gary, Indiana, Jackson began performing with his four brothers in the pop group the Jackson 5 when he was a child. Jackson went on to a solo career of astonishing worldwide success, delivering multiple No. 1 hit songs. He also starred in the 1983 film "Moonwalker" and released the best selling album of all-time, "Thriller," in 1982. He died on June 25, 2009 in Los Angeles, CA.	<p>Every sentence in this response contains factual information. You may be familiar enough with Michael Jackson to confidently know some factual information is accurate without further research such as he is widely regarded as the "King of Pop" or he was in the pop group the Jackson 5 when he was a child. You should focus your research on the important factual information (for example, he delivered multiple No. 1 hit songs) and the suspicious facts/claims (for example, the dates and locations).</p> <p>The Response is <b>Reasonably accurate</b> for two reasons:</p> <ul style="list-style-type: none"><li>• The movie "Moonwalker" actually came out in 1988. While verifiably incorrect, this fact is minor in the</li></ul>



		<p>context of answering the question "Who is Michael Jackson?"</p> <ul style="list-style-type: none"> <li>The claim "he was the most popular entertainer in the world in the early and mid-1980's" would be widely viewed as accurate by non-experts, but isn't an established fact.</li> </ul>
Tell me about the movie Moonwalker	Moonwalker is a 1983 American concert film directed by Colin Chilvers and Jerry Kramer. It stars Michael Jackson as he performs songs from his Bad album, and features performances by Joe Pesci, Sean Lennon, and Wesley Snipes. The film also includes a short film about Jackson's fictional alter ego, Moonwalker, who is on a quest to defeat a drug dealer named Mr. Big.	<p>Similar to the previous example, the Response incorrectly states the movie was released in 1983 instead of 1988. However, in this case the Prompt is specific to Moonwalker and the release year is an important fact about the movie.</p> <p>The Response is <b>Not accurate</b>.</p>
Tell me about the movie Moonwalker	The first humans to step on the moon were American astronauts Neil Armstrong and Buzz Aldrin as part of NASA's Apollo 11 mission on July 20, 1969.	<p>While the Response does not contain any inaccurate information on its own, it is an inaccurate response to the question in the Prompt.</p> <p>The Response is <b>Not accurate</b>.</p>
I have \$1000 to invest, what should invest it in?	Commodities such as gold and oil are a great investment, they tend to hold their value even when the economy is doing poorly.	<p>The Response makes a factual statement that "commodities as a great investment", but this is debatable and certainly would depend on the user's individual circumstances.</p> <p>The Response is <b>Questionably accurate</b></p> <p>If the Response was "one investment option is commodities" or "some people consider commodities to be a great investment, the Response to the Prompt would be a subjective statement and should be rated Completely accurate.</p>
Write a story about unicorns	A unicorn was walking in the forest when it came across a beautiful lake. The unicorn looked at its reflection in the water and realized that it was the only unicorn in the world. The unicorn was so sad that it cried, and its tears fell into the lake. The unicorn's tears turned the lake into a magical healing pool, and anyone who swam in the pool was cured of all their illnesses.	<p>The Response in the context of the Prompt is understood as a fictional story with no facts/claims.</p> <p>The Response is <b>not attempting to convey factual information</b>..</p>
When was Abraham Lincoln born?	He was born on February 12, 1809.	<p>The Response only contains one fact, which reliable sources will tell you is correct.</p> <p>The response is <b>Completely accurate</b>.</p>
How many pandas are currently in zoos, worldwide?	I don't have any information on that. I apologize for any inconvenience this may cause.	<p>This is a canned response, which is <b>not attempting to convey factual information</b>.</p>