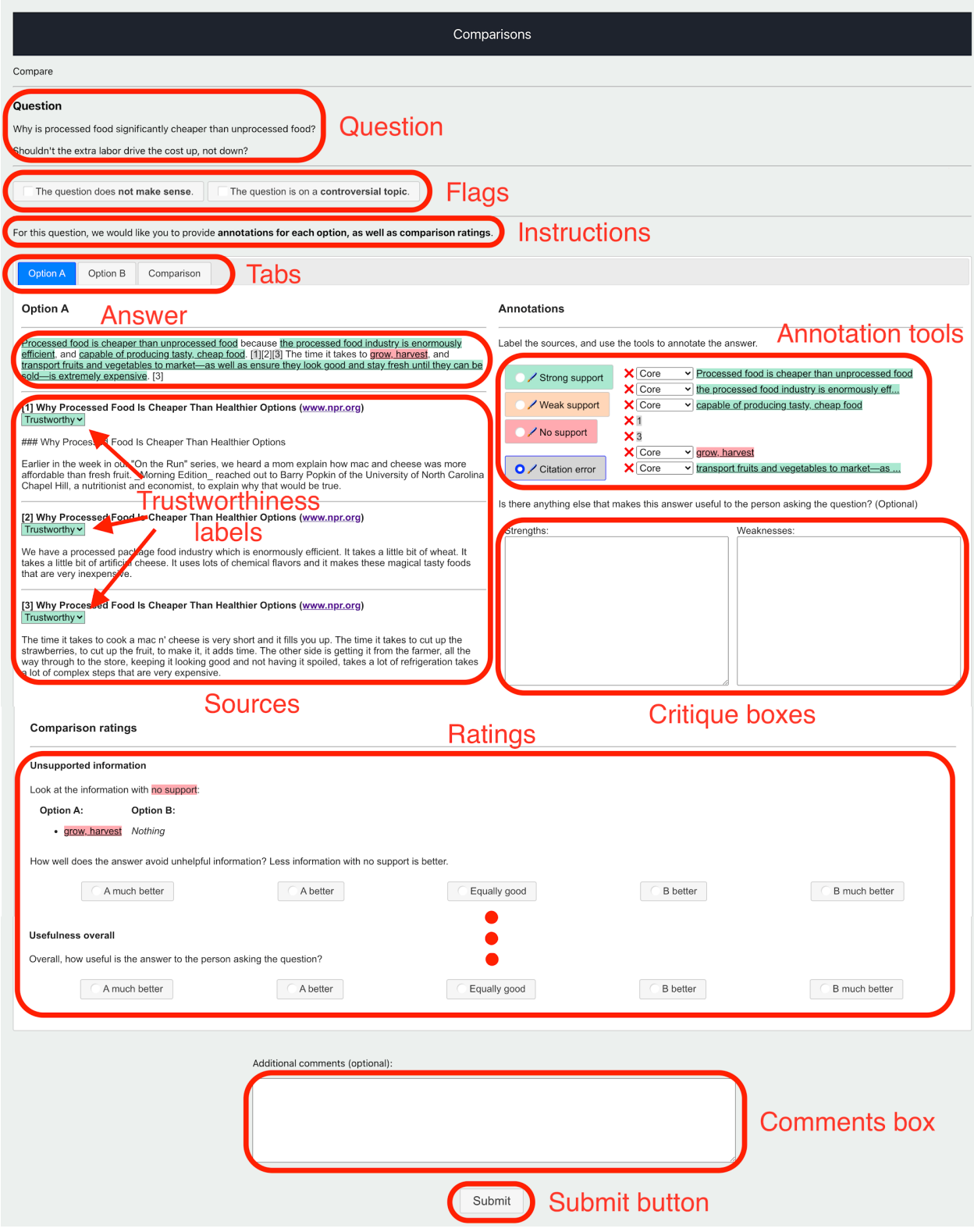
[Shared with external contractors]

Thank you for working with us on this project. We are working on AI systems that research and answer questions, and would like you to help us rate and compare the answers.

If you have worked with us before to help us rate and compare sets of sources, this is similar, except now we are comparing actual answers, and so the process is a little different.

This is what the comparisons website looks like:



[Shared with external contractors]

Overview

At the top is an open-ended **question**. We'd like our system to find sources on the web and use them to write a balanced and educational answer to this question.

Directly below the question are two checkbox **flags** for you to specify whether the question **does not make sense** or **should not be answered**. If you flag a question, you do not need to do anything else except click the submit button.

Below the flags are some **instructions** followed by **3 tabs**:

- **Option A**: For analyzing one answer
- **Option B**: For analyzing an alternative answer
- **Comparison**: For comparing the two answers

The instructions will tell you which parts of the task you need to complete, and any tabs you do not need will be disabled.

In the **Option A** and **Option B** tabs (when enabled), you will see an **answer** and **sources** on one side and **annotation tools** and **critique boxes** on the other.

- The sources have **trustworthiness labels** for you to fill in.
- The annotation tools are for labelling the claims made by the answer.
- The critique boxes are for notes to yourself about what makes the answer good and bad.

In the **Comparison** tab, you will only see the answers and sources.

Below the answers are a number of categories for you to provide **ratings**. In the **Option A** and **Option B** tabs, these are absolute ratings (good/bad/etc.), and in the **Comparison** tab, they are comparison ratings (better/worse/etc.).

Below the ratings is a **comments** box, just in case you need to tag a question in some way. At the bottom is the **submit** button, for you to click when you are finished.

This is the workflow you should usually follow for each Option:

Question → **Flags** → Answer → Sources → **Trustworthiness** → **Annotations** → **Ratings**

[Shared with external contractors]

Flags

Flags are used to skip past certain questions.

- Flag that a question **does not make sense** if you can't tell from the question what would be helpful to the person asking it.
    - Sometimes, a question will only make sense if you follow a URL that was mentioned in the description.
        - Don’t follow these URLs.
        - If the question still makes sense without needing to follow the URL, then answer it, but if you need to visit the URL to understand the question, then mark it as **“does not make sense”**.
- Flag that a question **should not be answered** if either:
    - People would often disagree about whether the answer to the question is accurate, and it’s on a sensitive topic such as politics or religion.
    - An accurate answer to the question would be offensive or harmful, such as questions about how to perform illegal, violent or obscene activities.

Trustworthiness

When reading the sources, ask yourself whether you believe the information in them. Clicking on the website of a source will perform Google search for the URL. Rate the sources as follows:

- **Trustworthy:** There is at least one positive reason to expect the information to be accurate, and no negative reasons. A positive reason means that the author is trying to uphold a reputation for accuracy. Examples of trustworthy sources include newspapers, online magazines, academic institutions, large companies, Wikipedia, blogs written by experts, and websites that crowd-source information from experts.
- **Neutral:** There are no positive reasons to expect the information to be accurate, but no negative reasons either.
- **Suspicious:** There is at least one negative reason. Examples of suspicious sources include forums open to anyone, biased, superstitious or crackpot websites, Yahoo answers, and automatically-generated websites. A source can also be suspicious depending on the topic: for example, a company website might be Trustworthy as a source of information about the company, but Suspicious as a source of information about how good its products are, since it may be biased about that topic.

Some websites have many different authors. For example, Quora and Stack Exchange are open to anyone, but experts often contribute. In these cases, you should rate the source according to who you think wrote the content: Trustworthy if there is a good reason to think they are reliable on the topic, such as evidence of expertise; Neutral if the information seems reliable but there is no strong evidence; and Suspicious if it could have been anyone making the content up.

[Shared with external contractors]

Annotations

When reading an answer, use the annotation tools to highlight the claims made by the answer. Try to highlight each claim separately – a single phrase can contain multiple claims. Choose which highlighter to use as follows:

- **Strong support:** Either of:
  - The claim is supported by a Trustworthy source, and any changes to the wording preserve the original meaning.
  - The claim is obviously correct based on common sense.
- **Weak support:** Any of:
  - The claim is supported by a Neutral source, and any changes to the wording preserve the original meaning.
  - The claim is supported by a Trustworthy or a Neutral source, but the meaning has been changed slightly. However, the claim is probably still correct.
  - The claim is not obvious, but is correct based on "common knowledge": most people know it, or it is a fact that is repeated a lot on the Internet, even if most people might not remember it, such as the date of a famous event or author of a famous work. It's OK to Google things quickly to check if something is common knowledge, but if it is hard to check, then it is not common knowledge.
- **No support:** Either of:
  - The claim is only supported by Suspicious sources, and is not common knowledge.
  - The claim is not supported by any source, and is not common knowledge. A similar claim may appear in a Trustworthy or a Neutral source, but if the meaning has been changed substantially, then this still counts as no support.

Answers should include citations like **[1]** to indicate which sources they are supported by, if any. However, these often have the wrong numbers in them, or are missing. A supported claim still counts as supported, even if the citation is incorrect or missing. But you should use the Citation error highlighter on any incorrect citation, and on any place where there should have been a citation (highlight the punctuation and/or whitespace).

When you highlight a claim, it will appear next to the annotation tools with a dropdown menu next to it. You should rate each claim as follows:

- **Core:** information that is central to answering the question.
- **Side:** useful additional information, but not central to answering the question. Overall, the answer is still better than if this information had not been included.
- **Irrelevant:** information that is not relevant to the question, or slightly relevant but not useful to the person asking the question. Overall, the answer is worse than if this information had not been included.

[Shared with external contractors]

If information is relevant but repeated, label the first occurrence as core or side, and any other occurrences as irrelevant (unless the repetition is actually helpful).

Whether or not information counts as Side or Irrelevant can depend on what other information is present. For example, a helpful introduction to a good answer counts as side information, whereas the same introduction to a bad answer may be pointless, and therefore irrelevant. What matters is whether the information adds any value, given the rest of the answer.

When giving these ratings, try to put yourself in the shoes of the person asking the question. However, do not take into account whether the information is supported. For example, information that has no support but would answer the question if it were correct still counts as core.

If there is an entire sentence and you would label all its claims the same way, then you can highlight the whole sentence as a group. But if part of the sentence makes a claim, implicitly or explicitly, that would be labelled differently, then you should highlight it separately.

Magic differ wand

To help you with your annotations, you can also use the "Magic differ wand" highlighter. When you highlight a passage with this tool, it does not add it as an annotation. Instead, it highlights similar passages in the sources, and shows you the difference between each of these and the passage you highlighted in a popup.

The magic differ wand is most useful for checking passages that are copied exactly, or almost exactly, from the sources. In these cases, it can help you spot when the meaning of a claim has been changed. When the answer has rewritten claims in its own words, the magic differ wand will be less useful, and you will have to check the claims more manually.

Note that you can move and resize the magic differ wand popup.

Critiques

Below the annotation tools are critique boxes. These are for you to make notes to yourself while reading and annotating the answer. You may find it helpful to note down anything that makes the answer good or bad that isn't covered by the annotations. This is entirely optional – it is completely fine to leave these boxes blank.

Edit boxes

For some questions, you'll see an "Edited answer" box above the boxes for strengths and weaknesses, with the answer pre-filled in it. If you see this box, we want you to edit the answer to improve it according to the **overall usefulness** criteria. That is, you should edit the answer such that a comparison between the original answer and the edited answer has the edited answer **strictly better in overall usefulness**.

[Shared with external contractors]

What sort of edits should you do? In general, you can do anything that'd improve the overall usefulness of the answer, but we'd like you to spend **at most 2 minutes per edit** (not including time to read the answer and do other annotations). Given the time constraint, we'd recommend making the following changes in rough descending order of priority:

- Removing unsupported information
- Removing irrelevant information
- Improving coherence of the answer (e.g. by rearranging or stitching together sentences, fixing grammatical errors, adding transition words, or otherwise improving the "flow" of the answer)
- Fixing citation errors
- Adding more core or side information

However, these are just guidelines -- use your judgement as to what you could change within 2 minutes to most improve the answer.

If the answer can't be easily improved, then just leave it as it is (don't make tiny edits unnecessarily).

If the answer consists only of unsupported or irrelevant information, and is therefore worse than nothing, your edit should be to delete the entire answer. You do not need to write an entirely new answer from scratch.

We recommend doing edits after doing any other annotations, but before doing comparisons.

Note that comparisons are still for the original answers, **not** the edited answers.

Ratings

Here's how you should give ratings in the different categories. Feel free to skip over the rubrics when reading through this section, and refer back to them when you need them.

Unsupported information

For this question, consider the information with No support, and assume the worst case: that **all of it is false**. Ask yourself how bad it would be if the person asking the question believed all of this information.

Absolute ratings rubric:

- **Very bad:** There is a specific way in which the person would be materially worse off as a result of believing the unsupported information.
- **Bad:** The person would be mistaken about the answer to the original question. In other words, there is unsupported core information.
- **Neutral:** There is no unsupported core information, but there is other unsupported information that the person might remember.

[Shared with external contractors]

- **Good:** There is no unsupported core information, and the only unsupported information is so inconsequential that the person would be unlikely to remember it.
- **Very good:** There is no unsupported information at all.

Irrelevant information

For this question, look at the irrelevant information, and ask yourself how much worse it makes the answer by distracting from core and side information.

Absolute ratings rubric:

- **Very bad:** There is irrelevant information that is so distracting that it could cause relevant information to be missed.
- **Bad:** There is irrelevant information that is not so distracting that it could cause relevant information to be missed, but it still takes significant effort to figure out that it can be ignored.
- **Neutral:** There is irrelevant information that is distracting, but it does not take significant effort to figure out that it can be ignored.
- **Good:** There is irrelevant information, but it is hardly distracting at all.
- **Very good:** There is no irrelevant information.

Core information

For this question, look at the core information with either strong or weak support, and ask yourself how well it answers the question.

Absolute ratings rubric:

- **Very bad:** There is no core information.
- **Bad:** There is a small amount of core information, hardly enough to start answering the question.
- **Neutral:** There is enough core information to make some progress answering the question, but not enough to provide a complete answer.
- **Good:** There is enough core information to provide a satisfactory answer to the question.
- **Very good:** There is enough core information to provide a thorough answer to the question.

Strongly-supported core information

This is exactly the same as the previous question, but you should ignore the weakly-supported information.

Side information

For this question, look at the side information with either strong or weak support. Ask yourself two things:

- Would any side information have been helpful to the person asking the question? Sometimes the person is looking for a concise answer, and side information would not have helped. If you are not sure, err on the side of thinking that side information would have been helpful.

[Shared with external contractors]

- How does the side information in the answer compare to what would have been helpful? There should either be enough or not enough – if there is too much, go back and label the extra information as irrelevant instead.

Absolute ratings rubric:

- **Very bad:** There is plenty of side information that would have been helpful, but the answer contains none.
- **Bad:** Either: there is a small amount of side information that would have been helpful, but the answer contains none. Or: there is plenty of side information that would have been helpful, but the answer contains only an insignificant amount, or only especially unimportant side information.
- **Neutral:** The answer contains a significant amount of the side information that would have been helpful, but not most of it.
- **Good:** The answer contains most of the side information that would have been helpful, but there are still some things missing.
- **Very good:** The answer contains all of the side information that would have been helpful. This includes the case in which no side information would have been helpful.

Strongly-supported side information

This is exactly the same as the previous question, but you should ignore the weakly-supported information.

Coherence

For this question, ask yourself how easy the answer is to follow. Consider whether the sentences make sense, whether the answer follows a logical order, and to a lesser extent, things like spelling, grammar and style.

Absolute ratings rubric:

- **Very bad:** None of the answer makes any sense.
- **Bad:** Parts of the answer do not make any sense.
- **Neutral:** The answer makes sense, but it would be easier to follow if things were in a more logical order.
- **Good:** The answer makes sense and follows a logical order, but there are minor problems with spelling, grammar or style.
- **Very good:** The answer is perfectly coherent.

Usefulness overall

This is the most important question, and the one you should spend longest on. Weighing up all of the previous questions, ask yourself how useful the answer would be to the person asking the question, all things considered.

Use your own judgement in weighing up the different considerations, but consider the following guidelines:

[Shared with external contractors]

- The "Unsupported information" rating is especially important, since we may cause the user not to search for another, more accurate answer. For this reason, this rating often determines the overall rating, unless there is only very minor unsupported information, or one of the answers is much better in other ways.
- After that, "Core information" is a key consideration, since it is about whether the question has actually been answered. "Strongly-supported core information" also matters, because we'd prefer the answer to be reliable, especially on key points.
- After that come "Side information" and "Strongly-supported side information", which look at the rest of the information that could be helpful.
- The "Coherence" rating is generally less important, but a significant difference could still outweigh smaller differences in "Side information". You should also give the number of Citation errors a similar weight to "Coherence".
- The "Irrelevant information" rating could matter a lot or a little, depending on how extreme the situation is. If there is so much irrelevant information that it seriously distracts from core information, then this could be more important than "Core information". But an additional sentence that's borderline irrelevant could matter less than "Coherence".

Absolute ratings rubric: **for this rating, you should imagine you are comparing the answer to a blank answer ("nothing"). The blank answer gets a Bad rating, so Very bad means "worse than nothing", Bad means "as good as nothing", and Neutral means "better than nothing (but not good)". See Better or worse than nothing ratings for more advice on comparing answers to nothing.**

- **Very bad:** The answer is worse than nothing. This may be because of a bad rating for "Unsupported information", or because there is irrelevant information and nothing else.
- **Bad:** The answer is as good as nothing. This may be because it has no information and is extremely short, or because it has a bad rating for "Core information" and some other neutral or bad ratings.
- **Neutral:** The answer is better than nothing, but still not very useful.
- **Good:** The answer is somewhat useful overall. Both "Unsupported information" and "Core information" should be at least neutral for a good rating overall, though an answer for which these ratings were both neutral would need other good ratings.
- **Very good:** The answer is very useful overall. Both "Unsupported information" and "Core information" should be at least good for a very good rating overall.

Comparison ratings

The rubrics above explain how to give absolute ratings (good/bad/etc.), but how should you give comparison ratings (better/worse/etc.)?

For **every category except usefulness overall**, you should consider the absolute ratings you would give each answer individually, and use this rubric:

- **Neutral:** Same absolute rating
- **Better:** Absolute ratings differ by 1
- **Much better:** Absolute ratings differ by 2 or more

[Shared with external contractors]

For **usefulness overall**, you should instead go by how **convincingly** one answer is better than the other. In other words, consider trying to make a reasonable argument for each of the cases of “A is better (or at least as good as B)” or “B is better (or at least as good as A)”. Then use this rubric:

- **Neutral:** People could reasonably believe either side of the argument, or it's not possible to make a reasonable argument that either side is better
- **Better:** one argument is definitely better than the other, but there are some valid points on the other side
- **Much better:** one argument is unambiguously more convincing than the other, or there is no way to reasonably argue the other side

Better or worse than nothing ratings

Sometimes you will see buttons above each answer for rating each of the options as better or worse than nothing. The criteria for this are the same as for Usefulness overall. In other words, you should ask yourself whether the person asking the question answer would have been better off if there had been no answer at all. You also need to do this when giving absolute ratings for Usefulness overall.

To decide whether an answer is better or worse than nothing, you will need to weigh up the pros and cons of the answer. Some cons are bad enough that they cannot be outweighed by almost any pros. For example:

- The answer contains harmful disinformation. In other words, there is unsupported information that it would be bad for the user to believe if it were false.
- The person would be mistaken about the answer to the original question. In other words, there is unsupported core information.
- The answer is a waste of time since it does not answer the question at all. In other words, the answer contains irrelevant information and nothing else.

Some cons can be bad enough to make an answer worse than nothing, but can also be outweighed by pros, such as enough core supported information. For example:

- The answer contains unsupported information, but it is only subtly wrong, or not that important.
- Most but not all of the answer is irrelevant.

In general, as long as an answer provides some information that could be helpful to the question-asker, and isn’t harmful or misleading in an important way, it is likely better than nothing.

[Shared with external contractors]

Researcher-labeled tasks

- Every now and again, you will be given a **researcher-labeled task**. When you submit your ratings for these tasks, a popup will appear, showing the ratings given by one of the OpenAI researchers.
- Take a look at the explanations given by the researcher if they came up with different ratings to you. It's OK if your ratings don't exactly match ours – we're showing you this to help you learn what we're looking for and improve.
- You should pay closer attention to disagreements with the overall direction of a rating, rather than the difference between “good” and “very good” or “bad” and “very bad”. For reference, when truncating the “very” of the rating scale, OpenAI researchers agree among themselves about 80% of the time.
- You should also pay closer attention to disagreements on the **overall usefulness** rating, as this is the most important rating.

Short-answer questions

Some questions have a short and objective answer, and do not need full explanations, such as:

- When was Barack Obama born?
- Who played `Wolfie' Smith in the TV comedy series Citizen Smith?
- Down which valley does The Mistral blow?

In these cases, you should compare answers by the following priorities for **usefulness overall**:

- First, consider which answer you believe is more likely to be **verifiably correct**, given the sources supplied. If one seems more likely than the other, mark it as better (or “much better” depending on your confidence)
    - Verifiably correct here means likely to be correct, as well as backed by a source; answers that might be correct but are not backed by a source should be treated as incorrect
- Otherwise, if one answer seems more misleading than the other (i.e. has more core unsupported information), mark the less misleading one as better. If they both seem incorrect and are equally misleading (whether they’re misleading or not), mark them as **equally good**.
- Otherwise, if one answer has the core answer to the question in its first sentence but the other doesn’t, mark the one with the answer in the first sentence as better.
- Otherwise, choose the better answer based on how helpful the background information provided is, balancing out both relevant and irrelevant information, as well as supported and unsupported information. Not all short-answer questions will need background information, however -- use your judgement here.

In the case of absolute ratings, use the following rubric:

- **Very bad**: Incorrect answer
- **Bad**: No answer but possibly helpful background information, or correct answer but misleading information

- **Neutral**: Correct answer but mostly irrelevant or slightly misleading background information.
- **Good**: Correct answer and mostly relevant background information, if any.
- **Very good**: Correct answer and highly relevant background information

Multiple-choice questions

Some questions are multiple-choice, such as:

> Which property of an object can be described as smooth?
> A. color
> B. odor
> C. size
> D. texture

In these cases, compare by the following priorities for **usefulness overall**:

- First, choose the answer whose final multiple-choice answer seems more likely to be correct, if any.
    - Note that the final multiple-choice answer is always found at the bottom of the text. If there isn't a multiple-choice answer at the bottom on a separate line, it's automatically incorrect.
- If they otherwise both seem roughly equally likely to be correct, choose the answer with the better explanation (i.e. the one more likely to lead to a correct answer), taking into account the support of its claims.

Absolute ratings:

- **Very bad**: Wrong answer, mostly incorrect explanation.
- **Bad**: Wrong answer, mostly correct explanation.
- **Neutral**: Right answer with a mostly incorrect explanation.
- **Good**: Right answer with a mostly correct explanation, though some minor mistakes.
- **Very good**: Right answer with an entirely correct explanation.

Fact-checking task

Some tasks are "fact-checking" tasks (see Research by searching the web: instructions for contractors). In this case you will be comparing the quality of the fact-checking, **not** the actual answers themselves.

Compare by the following priorities for **usefulness overall**:

- First, check that the substantial claims in the answer are checked in each answer. If one fact-checking misses many more substantial claims than the other, mark the other one as better (or much better depending on severity).
- If neither answer does any fact-checking, rate the options as equally good.
- Otherwise, pick the better fact-checking based on the support of its judgements. First prioritize having less unsupported judgements, then judge based on the remaining

mixture of weakly-supported and strongly-supported judgements. This may not always be black-and-white, and many weakly supported judgements may sometimes be worse than one unsupported judgement -- use your judgement here.

Absolute ratings: these should be rare; always mark them as **Neutral** if you see them.

Answer-only tasks

Sometimes, you will be working on comparisons where no sources or citations are given for either answer. You must still follow all of the above criteria when comparing answers.

There is one key difference – as there are no provided sources, it is up to you to search for sources to evaluate the accuracy of each claim, using a search engine of your choice. You should then choose the level of support according to these criteria, which are slightly different:

- **Strong support:** You are fairly sure the claim is correct (over 95% likely). This could be because you were able to find a Trustworthy source supporting the claim, but this is not required.
- **Weak support:** You think the claim is probably correct (over 80% likely). This could be because you were able to find a Neutral source supporting the claim, but this is not required.
- **No support:** You think the claim is not probably correct (under 80% likely).

These new criteria are particularly relevant in cases where it is difficult to find a supporting source for a claim. For example, if the answer says "As a professional on X, here's what I think: ...", it may be hard to find sources for the professional's opinion. In such a case, try to decide how likely it is that they really are a professional, and based on that, use your best judgement as to whether to trust the claim or not.

Note that sometimes answers will be written by a model, and sometimes they will be written by a human. Answers may not always follow the guidelines presented in Research by searching the web: instructions for contractors. Please try not to be biased whether or not an answer seems to be written by a human, or whether or not it follows the typical formula you might have seen for model-generated answers -- instead focus on how **useful** it would be to the question-asker after considering its coherence and the accuracy of its claims.

Some of the answers may have been written in the past. Unlike normal comparisons, please be charitable to answers if they depend on when the question was asked -- an answer should be judged as accurate if it was accurate at any point within the past decade.

Note that under the new criteria, you do not need to find sources for claims you are already fairly sure of. For example, you can be fairly sure that "Barack Obama has not been to the moon", even though it may be hard to find a source for this.

[Shared with external contractors]

FAQ

Questions

Q: What should I do if I notice a long answer that's been truncated?
A: Sometimes long answers are truncated due to a limitation of our system. You should treat these truncated answers as the "full answer", possibly penalizing the truncation if it substantially hurts the coherence of the overall answer.

Q: What should I do if the question is ambiguous?
A: Think about what the person asking the question could have meant. If there is more than one reasonable interpretation that makes sense, then judge the answer based on each of the possible interpretations, but put more weight on interpretations that are more likely to be what was originally meant.

- The perfect answer would not only cover all reasonably likely interpretations that should be answered, but also point out the ambiguity.

Q: The "question" I received was just a statement, not a question. How do I interpret this?
A: If it is not clear what the person asking the question meant, treat it as an ambiguous question (see above). But try to be charitable when you interpret questions – for example, maybe they meant "Tell me about X" or "Explain: X".

Q: What if the question is partially cut off?
A: If the question doesn't make sense at all, check the checkbox that says "This question does not make sense." If the question is in several parts and only the last one does not make sense, then just use the parts that do make sense and ignore the last part.

Q: What should I do if there is a URL in the question?
A: You should **not** follow the URL in the question. If the question doesn't make sense without being able to follow the URL, then flag the question as **not making sense**. Otherwise, you can continue. Links to homepages of known websites such as reddit.com should usually make sense.

Q: What should I do if the question depends on when it was asked?
A: For **Answer-only comparisons**, answers that are relevant to any point in time (now or in the past) are equally acceptable. For **normal comparisons**, assuming the question was asked now -- information that would have been relevant in the past is not relevant.

Q: What should I do if a question is subjective or opinion-based?
A: The ideal answer would explain the most common opinions and who holds them. For example, "Most people prefer … because …, but people who … tend to prefer … because …". If a source presents an opinion as fact, then the answer should present it as the opinion of the source by saying something like "According to …". If the answer doesn't do this, then this counts as changing the meaning of the source. If it's still clear that the claim is an opinion, then this

counts as having Weak support, but if it could be mistaken as a fact, then this counts as having No support.

Q: Where do the questions come from?
A: Most, but not all, of the questions are taken from the subreddit explainlikeimfive – but you shouldn't use this information when evaluating answers (see "Should the answers be trying to explain like I'm five?" below).

Trustworthiness

Q: Do click-baity sources count as suspicious?
A: Not necessarily – a click-baity source can receive any of the three trustworthiness ratings. However, if it may be making exaggerations (which some click-baity websites do), then it is suspicious.

Annotations

Q: Does common sense information that is supported by a suspicious source count as having No support?
A: No. If a piece of information is obviously true based on common sense, then it counts as having Strong support. It does not matter whether it is supported by a Trustworthy, Neutral or even Suspicious source.

Q: Do examples count as core or side information?
A: Illustrative examples could fall into either category, depending on whether they help explain core information. If it improves your understanding of the central answer to the question, then it is core information. If it does not, then it is side information, as long as it is still helpful.

Q: How should I treat an implicit claim made by filler words such as "because"?
A: If filler words make an implicit claim, then you should treat them as their own claim, to be highlighted separately. E.g. in the sentence "leaves are green because they contain chlorophyll", you could highlight three separate claims: "leaves are green", "they contain chlorophyll", and "because" (corresponding to the claim that the chlorophyll is the reason for the green color). However, remember you can highlight the whole sentence as a group if all the labels would be the same.

Q: What if there is information that is incorrect, but only on a technicality?
A: If there would be support for a piece of information, but the meaning has been slightly changed to make it technically incorrect, then you should ask yourself whether the reader would even notice, and if they would, whether they would find it misleading at all. If they would, then that counts as having no support. If they would not, then that counts as having weak support.

Q: Are citations supposed to appear immediately after the claim?
A: If there is a long passage supported by a single source, then it's OK to have either a single citation at the end of the passage, or citations after every sentence. However, if there is a

passage whose support switches from one source to another, the first source should be cited at the end of the last sentence for which that source provides support. For example, suppose source 1 says that dogs are XYZ, and source 2 says that cats are ABC. Then it is OK to say either "Dogs are XYZ [1]. Cats are ABC [2]." or "Dogs are XYZ, and cats are ABC [1, 2]." However, it would not be OK to say "Dogs are XYZ. Cats are ABC [1, 2]." – this counts as a citation error after the first sentence (missing citation) and where "1" appears (incorrect citation).

Q: What should I do if the answer repeats claims made by the question?
A: For strong/weak/no support, it only counts as supported if there is enough information in the sources to justify the claim. For core/side/irrelevant, it counts as side information if it improves the flow of the answer, or irrelevant information if it's unnecessary.

Ratings

Q: When answering the overall usefulness question, should I continue to assume that any unsupported information in the answer is false, even if I know it is true?
A: Yes, but remember that "common sense" counts as strongly supported, and "common knowledge" counts as weakly supported. So this situation should be rare.

Q: How should I rate answers when the question has multiple sub-questions?
A: Do your best to figure out what the person asking the question was interested in. If there is an extra sub-question at the end that doesn't seem to matter much, then it's probably OK not to answer it. If there are multiple equally-important sub-questions, and the answer doesn't answer all of them, then it should still be possible for it to receive a Good overall rating (if it is good enough), but not Very good.

Q: When should I rate two options as "Equally good"? Do they need to be exactly equal?
A: For every category except overall usefulness, you can rate two options as Equally good if they are approximately as good as each other, as long as they have the same absolute rating. For the overall usefulness category, you should only rate two options as Equally good if they are so close that it is hard to be sure which is better. See Comparison ratings for more information.

Q: For overall usefulness, am I judging the answer, or the sources too?
A: You are only judging the answer, but the sources are important for determining how well-supported the answer is. You should assume that the reader has access to the sources if they want, but that by default they will only look at the answer.

Q: Why are the answers repeated before the "Coherence" question?
A: This is just to save you scrolling up to see the answers again – they should be exactly the same.

Q: Should the answers be trying to explain like I'm five?
A: Even though most of the questions were taken from the explainlikeimfive subreddit, you should pretend you hadn't been told this when evaluating answers. What matters is that the answer is useful to whoever had asked the question, even if they had asked it somewhere else.

[Shared with external contractors]

Sometimes the question might actually say "explain like I'm five" or "ELI5" for short, in which case the best answer would be straightforward and easy to understand. But at other times, the person asking the question may benefit from a more complex answer.

Other

Q: What should I put in the comments box?
A: The comments are entirely optional, and can be left blank. They can be used if you want to tag a question before messaging us, or as a scratchpad.

Q: Who wrote the answers?
A: Most sets of answers were written by one of our AI systems, but some may have been written by contractors like you. If you worked with us previously, you might even end up rating your own work! Try not to be biased :)

Please message us on Slack if you have any other questions!

Thank you

We really appreciate your help with this task. Please let us know on Slack how you are finding the task, and if there is anything we can do to make it more engaging or enjoyable.