Annotation Guidelines

1. Procedure

In this task, I will show you question-answer pairs from Reddit, which is an online question-answer forum. I ask you to evaluate the *helpfulness* of each answer on the 1-5 scale, according to the following instructions in the next section. We use a five-point Likert scale ranging from 1 (not helpful) to 5 (very helpful).

Helpfulness is a multifaceted concept and depends on the context of every question, which makes it difficult to define helpfulness in an objective manner. Thus, it is important to follow the given guidelines while annotating to produce consistent judgments. I will suggest five factors (**relevance**, **explanation**, **informativeness**, **entertainment**, and **readability**) that constitute the concept of helpfulness and examples corresponding to each factor.

When you rate each answer, **you will assign the score ranging from 1 to 5 for each evaluation item**. The final output is **the average of the five evaluation items rounded to the nearest integer**, which is helpfulness of the answer. Keeping this in mind, I introduce the five evaluation criteria as follows:

2. Measures

In this annotation task, you will evaluate the helpfulness of each answer based on the following evaluation items.

2.1 Relevance

The answerer should give a proper answer, which is closely linked to the given question. The answer should not create confusion or ambiguity. You should also deduct points if the answer looks obviously false.

Use the following 1-5 scale: 1 (very bad) – 2 (somewhat bad) – 3 (neutral) – 4 (somewhat good) – 5 (very good)

Example 1:

> The examples below would be a response that has **no relevance to the original question**, such as repeating the question, ads or solicitation for products not asked for, or discussing a topic other than the original question.
>
> **Q.** How far is the Sun from Earth?
> **A.** Jupiter is the largest planet in the solar system.
> **A.** Come play at the Sun Casino!
> **A.** The Sun is not in our solar system.

Example 2:

Q. When drunk what's the most embarrassing thing you've done?
A. That is classified information, and I would probably be on a FBI watchlist if I disclosed that.

Comment: The answerer refuses to give a proper answer and says something that has no relevance to the question. Thus, I will give the score of 1 in relevance.

Example 3:

Q. What is something you started doing recently and advise everyone to do it too?
A. Switch to decaf coffee after midday.

Comment: The answer is relevant to the question so I will give 5 points for relevance.

2.2 Explanation

A good answer should be backed by at least one proper **reason** or **description**.

Use the following 1-5 scale: 1 (very bad) – 2 (somewhat bad) – 3 (neutral) – 4 (somewhat good) – 5 (very good)

Example 1:

Q. [Falcon and Winter Solider] Would Frank Castle have been a better choice than John Walker?
A. Castle is a psycho. He would be off the charts worse.

Comment: The answerer mentions that "Castle is a psycho," but that does not sound like a proper reason so I will probably give 2 points in explanation.

Example 2:

Q. Being careful to avoid spoilers, what is the BEST ending to a movie, book, or series that you've come across?
A. Fight Club by far has the absolute best ending of a movie or book but for different reasons; the movie depicts the woes of consumerism while revelling in its inherently capitalist medium and creates an ambient and perfect movie with strong characters. The book offers a deeper philosophical look into toxic masculinity and the temptation of nihilism in young adulthood/teens. The absolute perfect mindfuck I promise, it will have you guessing if you don't know it. Greatest book or movie of all time.

Comment: The answer is supported by several different reasons so I will give 5 points for explanation.

Example 3:

Q. How far is the Sun from Earth?
A. 92 million miles.

Comment: The answer does not give any further details or resources as to where the answer can be found. I will give the score of 1 due to the lack of explanation.

2.3 Informativeness

Helpful answers tend to be useful in practice and contain new/creative information. They may also provide resources or details from credible sources.

Use the following 1-5 scale: 1 (very bad) – 2 (somewhat bad) – 3 (neutral) – 4 (somewhat good) – 5 (very good)

Example:

Q. What's an inappropriate question to ask during a wedding?
A. Asking the bride if she wants to see a 1 eyed trouser snake.

Comment: The answerer is not serious enough and says something too obvious, which is not helpful. I will probably give the score of 1 or 2 in informativeness.

Example:

Q. How far is the Sun from Earth?
A. The distance between the Sun and Earth is roughly 90 millions miles. The distance is constantly varying due to the elliptical orbit of the Earth about the Sun. That being said, light takes about 8 minutes to travel from the Sun to Earth. Knowing light travels at a speed of about 300 million meters per second, we can multiply that by 60 (seconds in a minute), then multiply that by 8 (minutes to travel). This gives us 144,000,000,000 which we can divide by 1,609 to give us miles instead of meters. The NASA website has great resources for such questions and calculations, I would recommend visiting the site.

Comment: The answer provides detailed information and resources where a reader can seek additional information. I will definitely give the score of 5 in informativeness.

2.4 Entertainment

It has been found that the use of humor is positively associated with review helpfulness (Schindler & Bickart, 2012). Note that the value of this item could vary from person to person.

Use the following 1-5 scale: 1 (very bad) – 2 (somewhat bad) – 3 (neutral) – 4 (somewhat good) – 5 (very good)

Example:

Q. What would happen if the Hoover Dam broke?
A. We would need Hoover vacuums.

Comment: Even though the answer is neither practical nor informative, it is entertaining so we may give credit to it. I will give 4 points for entertainment.

2.5 Readability

You may deduct a point if the answer is way too long or hard to understand. You may also deduct a point if the grammar is bad because bad grammar and spelling could affect our machine learning process.

Use the following 1-5 scale: 1 (very bad) – 2 (somewhat bad) – 3 (neutral) – 4 (somewhat good) – 5 (very good)

Example:

> **Q.** What one thing disgusts or angers you to your core?
>
> **A.** This is a super petty one but: I watch a lot of sponge rinsing ASMR. It's basically just videos where people rinse a soapy sponge. But the problem is that in at least 80% of videos, they don't actually finish rinsing the sponge. Like at the end there's still soap in it. And I always thought it was a universal thing to find it unsatisfying to see something left unfinished and incomplete. Like sometimes I've seen random YouTube videos that are supposed to be unsatisfying and they're always something being done unevenly or done partially and left incomplete. Like imagine you see a video of someone cleaning a window and they clean 90% of it but leave a dirty spot. That's unsatisfying and frustrating, right? Doesn't everyone feel that way? But apparently I'm the only one in the sponge rinsing community who cares if the sponges are fully rinsed, because in most videos they're not and those videos get just as many likes as the ones that actually finish rinsing and never have any dislikes. And just, how? I honestly find it frustrating and stressful when the rinse is left incomplete. And it actually angers me that nobody else cares at all, that even though you'd think it's universal to find it unsatisfying to see something just partially done and left unfinished and incomplete, nobody else cares. I mean the whole point of these ASMR videos is to be satisfying, they often have "satisfying" in the title, but it's exact opposite of satisfying to leave something incomplete! How does no one else see that? So since no one besides me cares, the people who make the videos don't bother to make sure the sponges are fully rinsed. Sometimes they will be, but in most videos they're not. And I really, really love the sponge rinsing ASMR videos where the sponges actually are fully rinsed, but I have to search through so many unsatisfying, frustrating videos where the rinse is only partially done in order to find them. It seriously makes me mad that I'm the only one who cares whether the rinse is actually finished or if the person just does a half assed job and only partially rinses the sponge and leaves soap in it at the end.
>
> Comment: TL;DR and I will give the score of 1 in readability.

Based on the five factors shown above, your task is to rate the helpfulness of each answer.

3. Annotation Examples

In this section, I will demonstrate how to annotate step-by-step, using the above guidelines and sample question-answer pairs. Note that the rounded rating is the final output – *helpfulness*. When you annotate the data, you can first create five columns and then fill in the scores for the

five evaluation items. Calculate the average rating and round the results at once by using Excel, Python, R, and so on.

Example 1

Q. What part/organ in the human body does more than some people think?
A. I guess the obvious one is the appendix. Formerly considered useless, it is now thought to be the factory and reservoir of gut flora: the bacteria that help us process nutrients from food.

Relevance	Explanation	Informativeness	Entertainment	Readability	Average Rating	Rounded Rating
5	5	5	1	5	4.2	**4**

Comment: The score of helpfulness for this answer is 4.
Relevance: The answer is related to the question.
Explanation: The answerer gave a proper reason.
Informativeness: The answer gives us new information about the function of the appendix.
Entertainment: There is no attempt to make the answer entertaining.
Readability: The answer is concise and easy to understand.

Example 2

Q. What is the best ice cream flavour?
A. Vanilla because you can't go wrong with it

Relevance	Explanation	Informativeness	Entertainment	Readability	Average Rating	Rounded Rating
5	3	1	1	5	3	**3**

Comment: The score of helpfulness for this answer is 3.
Relevance: The answerer mentioned the type of ice cream.
Explanation: At least, the answerer gave a reason for the answer even though it does not sound very strong.
Informativeness: The answer does not sound very interesting or informative.
Entertainment: There is no attempt to make the answer entertaining.
Readability: The answer is concise.

Example 3

Q. What are some things that would cause you to end a friendship for good?
A. I see them eating ketchup on scrambled eggs, it's over.

Relevance	Explanation	Informativeness	Entertainment	Readability	Average Rating	Rounded Rating
4	2	1	1	5	2.6	**3**

Comment: The score of helpfulness for this answer is 3.

Relevance: The answer is about the end of a friendship.

Explanation: The answerer does not give a proper reason.

Informativeness: The answer is not informative.

Entertainment: Some people may find this answer funny, but it is not funny enough for me.

Readability: The answer is concise.

References

Schindler, R. M., & Bickart, B. (2012). Perceived helpfulness of online consumer reviews: The role of message content and style. *Journal of Consumer Behaviour*, *11*(3), 234-243.