<center>**Team Annotation Guidelines**</center>

Introduction

The task of annotating responses based on their quality is often very involved. Because questions can vary largely in their meaning, useful or helpful replies may take many different forms. As a consequence, an understanding of context (particularly in relation to the question) is critical when attempting to annotate the quality of responses. When annotating responses for quality, it is often important to consider multiple attributes of a response in order to create an accurate judgement or rating.

Below we have included an annotation guideline detailing important qualities present in responses which may affect their quality. This list of attributes is not comprehensive, but provides a strong introduction toward understanding examples of quality responses. Each quality listed below also includes two examples of responses (high quality vs. low quality) along with a suggested ranking. Ranking suggestions listed below are either a single value or are accompanied by a "+" or "-" to signify that the rating may vary based on a specific response.

General Guidelines for Annotation

1. Length of Response

The length of a response often dictates its quality. Long responses typically provide more transformative content regarding the original question (meaning that the response adds new knowledge rather than simply restating the question) and thus commonly receive higher ratings. Short responses which do not provide additional understanding, new knowledge, or an answer to the question are rated lower.

Question :	"How do you outsmart your instincts that have evolved over millions of years?"
Response :	"What do you mean by outsmart?"
Rating :	2

Question :	"What is the dumbest thing you believed as a kid?"
Response :	"Airplane fuel is rainbow colored. I think I got it from looking at gas station colours for each type of fuel. My reasoning was that because an airplane was so big, it would need some sort of 'super gas' to be able to fly"
Rating :	4+

2. Responses can be On or Off-Topic

The content provided within a response also works to dictate its quality. If the content provide by the user is relevant and related to the topic of the question, then its quality is often perceived as better. Responses which are off-topic or shift to a different topic midway through a response are typically given lower ratings.

Question :	"A bottle is half and half, is it half empty or half full and why do you think that?"
Response :	"I'm half drunk"
Rating :	1

Question :	"Any advice for someone who wants to start creating stories?"
Response :	"During the process if creating characters, really get to know them. An example that has really stuck with me is that you want to be able to know what your character would have in their junk drawer."
Rating :	4+

3. Partial Answers

The quality of an individual response is largely based on the type of question asked. Below is a list of each of the common question types and what kind of content they expect as a response.

Question Type	Requirements
"What" Questions	An Example or List of Items/Instances
"Who" Questions	Person or Thing
"When" Questions	Time Frame and Typically an Instance
"Where" Questions	Require a Location
"Why" Questions	Explanation
"How" Questions	Instructions

Responses which do not provide content covering the requirements for a given type of question are given lower ratings as they only partially answer the question. One the other hand, responses which fulfill the requirements of a question and provide additional reasoning are often given higher ratings on account of being more detailed.

Question :	"Whats the best place on the internet to learn random interesting things?"
Response :	"Daily dose of internet on youtube"
Rating :	3

Question :	"People who did not sleep for more than 24hrs, what was it for?"
Response :	"A wisdom tooth was acting up. The pain kept me awake for 3 days."
Rating :	4+

4. Wellformedness of Responses

The quality of writing within a response has some effect on the overall rating of a response. The general rule of thumb is that as long as the response is understandable and successfully conveys its meaning, it can be given a higher rating. Answers containing numerous abbreviations or misspellings may result in a more confusing response and consequently receive a lower rating on average.

Question :	"What video game achievement are you most proud of?"
Response :	"I hit Diamond 4 in LoL S8."
Rating :	2-

Question :	"What will our society's fossil record look like?"
Response :	"Anthropocene: A layer of plastic in the sediments, with all the signs of global warming and high carbon in the atmosphere."
Rating :	5

5. Offense, Harassing, or Illegal Responses

Because responses to questions may arise from many different platforms and from a variety of different people, some responses may approach the question with a negative sentiment. For example, some responses may purposely harass or intend to offend a user while other responses may suggest dangerous or illegal activities. These types of responses typically receive lower ratings.

Question :	"What is something people make WAY too much of a deal about?"
Response :	"F***ing cow milk, y'all nasty for that."
Rating :	1

Question :	"What's that thing which you cannot possibly do sober but definitely try it when you're drunk?"
Response :	"go through a mcdonald's drive-thru in a battery operated toy vehicle."
Rating :	2-

6. Sarcastic, Joking, or Misinformative Responses

Once again, because the answers to questions may come from a variety of users and sources, responses may purposely be unhelpful or deceptive. Responses to questions which take the form of a joke or a sarcastic reply commonly receive lower ratings on account of their irrelevancy. Questions which purposely provide misinformation regarding a topic also receive lower rating. Judging for misinformation in a response can often be difficult, however, as the response to the question's topic may not be common knowledge for annotators.

Question :	"If you strive to fail, and succeed, which do you do?"
Response :	"You've succeeded in failing"
Rating :	2

Question :	"How did you earn your first million?"
Response :	"Shares in GameStop! (Just for clarification it hasn't happened yet but give it a week and we should be good 😉)"
Rating :	2+

7. Negated or Guessed Responses

Some user generated responses may not come from personal knowledge or experiences. Instead, these answers are often speculative and unhelpful to the original question. Such guessed responses are often given lower ratings. Other responses, in which a user provides transformative content, but negates their statement (saying that they are unsure or not an expert, for example) also receive lower ratings on account of the responding user's lower confidence in their answer.

Question :	"What is the most Average movie you can think of?"
Response :	"The Notebook. Never saw it but its what came to mind."
Rating :	3-

8. Responses Containing Links to External Sources

In many cases, external information or understanding can be helpful when responding to a question. When considering annotating responses based on their quality, however, annotators are only able to consider the text provided to them. As a consequence, short responses which only contain a link are considered unhelpful and given low ratings. Not all responses containing links to external sources or websites are bad — If the response also contains clarifying information or an understanding of what the link is meant to convey, then they can be awarded a higher rating.

Question : "What was something you laughed at that was morally wrong to laugh at?"

Response : "[this](https://youtu.be/yhKZCy41g5w)"

Rating : 1

Question : "What song would be played the most if a deadly apocalyptic event occured worldwide?"

Response : "A piece rather than a song, but the [second movement of Beethoven's 7th symphony](https://youtu.be/mgHxmAsINDk) has major apocalypse vibes for me."

Rating : 4