SI 630
Winter 2022
Group annotation guidelines
March 9, 2022

## Scoring Rubrics

In general, an ideal "helpful" answer is comprehensive and precise. It should sincerely answer to what the questioner asked, provide sufficient information to support the statement, and always use appropriate language.

For annotating the helpfulness of the answer, a scale from 1 to 5 will be applied. Relevance to the question is the prioritized criteria. Base scores are assigned to answers based on the judgment of relevance. We define 3 levels of relevance from low to high: absurd — acceptable — reasonable, which correspond to 1, 3, 5 points of base score respectively. Language appropriateness and information quality are then evaluated. Any violation of the rules will result in the deduction of corresponding points from base scores. If the final score is decimal, it should be rounded up to the nearest integer.

In extreme cases, an answer could be assigned a negative score because of too much violation. Since the lowest score we could annotate is 1 point, annotators should assign 1 point to those answers. For the same reason, annotators could skip the evaluation and deduction when the base score is 1 point.

The detailed criteria for each aspect are as follows.

**Relevance (reasonable — acceptable — absurd)**

- Reasonable (5 points): The answer attempts to sincerely address the question, and is consistent with the question context.

- Acceptable (3 points): The answer attempts to provide some related information within the same topic of the question, but not directly answers the question. This relevance level also includes answers from an unusual but somehow logicly acceptable perspective.

- Absurd (1 point): The answerer twists or misinterprets the question, and gives an irrelevant and/or unreasonable answer.

**Language appropriateness**

- Readability (1 point): Readers do not have to take extra effort to get value from reading an answer. The answer should avoid exclusively wording. If the answer is in length, a good organization and/or formatting will be expected. In addition, the answer should be coherent, which makes it easy for readers to understand the whole paragraph. If annotators find that the answer is hard to follow and/or unreadable, it will result in a 0.5 - 1 point deduction.

- Community Friendliness (0.5 point): A community-friendly answer is in proper wording and shows respect to others. If annotators find that an answer is anti-social, verbally abusing, and/or discriminatory, 0.5 point at most will be deducted from the total score.

**Information Quality**

- Explicit (1 point): Helpfulness is more than giving a brief answer. The response is also expected to provide sufficient information to clarify and support the statement. Supporting information includes further description, detailed explanation, examples and external resources. If the response only contains a short answer without any information in detail, 1 point to the maximum will be deducted from the total score.

- Comprehensive (0.5 point): A helpful answer should be generalized and reusable. Despite answering questions about personal experiences, solitary narratives of individual cases are discouraged and will lead to a 0.1 - 0.5 point deduction from the total score.

- Constructive (0.5 point): The answer to questions requiring specific solutions to an issue should provide feasible approaches rather than vague judgment and opinion. Violation will lead to a 0.1 - 0.5 point deduction from the total score.

- Special Requirement (0.5 point): A helpful answer should also meet the specific requirements as the question describes. For example, if the questioner specifies 'ELI5' in the question, which means explaining like I'm 5, the answer therefore is supposed to be as simple and straightforward as possible. Failure to observe special requirements will result in a 0.5-point deduction from the total score.

## Story/experience-related questions score categories

For questions asking for responses of stories, personal experiences or imagination (for example, the response to the question type "What will you do if …"), still, relevance is the most important criteria. If the response directly or indirectly describes the question prompt, its score should at least be 3. However, since responses to this type of questions do not necessary to satisfy many of the above rubrics, (i.e., requirements about explicitity ). The score categories are presented below:
5: Very detailed and credible descriptions of the story/experience (5 to 8 sentences would be suffice).
4: Some descriptions and explanations about the story/experience (3 to 5 sentences would be suffice).
3: Simply describes the story/expereince in one or two sentence.
2: Off-topic stroy/experience with some descriptions and explanations or it is relatively hard to understand the response.
1: Totally off-topic, or the question prompt is not answered.
It should be noticed that for these types of responses, whether it is helpful is very subjective. As a result, we believe the line between 4 points and 5 points is not very clear. When it is hard to decide whether the response is 4 or 5, please follow the feeling although it sounds very

irresponsible. If it is a very compelling story with fewer sentences, it can be scored as 5. If the story is very long and detailed, but lack of interests, it can be scored as 4.

## General Score Categories & Examples

Based on the scoring rubrics above, we provide specific descriptions of each score category.

**5 – Very helpful**

If the response satisfies **all of the 5 criteria**, it should be assigned to 5 points as "Very helpful".

- Respond directly to the topic instead of changing subjects or making fun. It is useful to provide some background information, but it should be limited to one to two sentences in order not to distract the main focus.

- The response would be not very helpful if it is too specific that only helps a certain group of people under certain circumstances. People should be able to generalize the idea from the response. It is also very useful to cover different cases with conditions. If the response considers these conditions, it satisfies this criterion.

- With reasonable explanation. People want to know the reason and logic behind the idea, so a very helpful response should include the thinking process. Also, it is important to make the response credible instead of wrong or incorrect.
    - If the response relates to personal opinion or individual experience, it should provide enough information and examples to prove the credibility.
    - If the response relates to facts, it should provide reliable sources or have citations.

- Clear and easy to read. There is no difficulty to read and understand the contents. Few small grammar or spelling mistakes (1 to 2) can be ignored.

- Show respect to the question. Do not attack or harass other people.

**4 – Somewhat helpful**

If the response satisfies **some of the 5-point** criteria and satisfies **all 3-point criteria**, it should be assigned to 4 points.

**3 – Neither helpful nor unhelpful**

If the response satisfies all these two criteria, its score should be 3 as "Neither helpful nor unhelpful".

- It responds to the question no matter directly or indirectly. In this category, the response at least is related to the question. The score cannot exceed 3 if it does not respond to the question.

- The evidence and explanation are not credible or reasonable enough to support the response contents.

**2 – Somewhat unhelpful**

If the response does not satisfy all of these 1-point four criteria and does not respond to the question, it should be annotated as 2 – "Somewhat unhelpful".

**1 – Not helpful**

If the response satisfies **all these 4 criteria**, its score should be 1 as not helpful.

- The grammar and spellings are so poor that these mistakes would negatively influence other people to understand.
    - It is common for people to have typos when they are using a mobile phone. Also, it is acceptable for non-native English speakers to make mistakes about what they say and write. If these mistakes do not affect the main contents about the response, it should be fine. However, if these mistakes make other people have trouble with understanding the response, this criterion should be satisfied. For example, a few misspellings (less than 10) can be dismissed. More than that would distract people's attention from the contents of the response.

- Too short. The response should have enough contents like examples and explanations that show how they reach the answer.
    - For example, for yes/no questions, the simple one word "Yes/No" is not helpful even if it is the fact because people need to know the reason.
    - One paragraph (3 to 8 sentences) is expected as a sufficient response.

- Without explanation. People generally want to know the rationale for the response to know whether it is reasonable or logical so that they could trust the answer. Simply personal opinions or emotions without explanation do not help in most cases.

- Show no respect. Except for anecdotes for stories, people want to hear the response from people who are willing to respond. As a result, it is very disrespectful for the response that attacks the question or other people.

## Miscellaneous

For question texts that are too sophisticated to understand or without context, the response score should be 3 because we do not know the true meaning of the question.

- For example, the question text is simply "How can I ignore them?" Since we do not know what "them" refer to, the response becomes meaningless. As a result, we cannot judge whether the response is helpful or not.

For the response that lack of explanation but common people seem to understand its meaning behind the wording based on conscious, its score can also be 3.