Evaluation Guideline for Annotated Question-Answer Pairs

Each evaluator receives an individual file that contains approximately 90 question-answer pairs with the paragraph. Each question-answer pair should be evaluated individually according to the scale system from 1 to 4.

Scale 1:

The question or answer is not written in correct English, and it is difficult or impossible to understand the meaning.

The question or answer is composed of severe semantic or syntactic errors.

The question proposed is not relevant to the topic of climate change. This includes questions that are too abstract or general, or impossible to answer with the information given in the text.

The proposed answer did not answer the question.

Scale 2:

The question or answer contains grammatical errors, and it takes some time to understand it.

The proposed question is relevant but unclear or difficult to understand.

The proposed answer only answers the question partially.

Scale 3:

The proposed question is understandable without investing too much time, but it contains syntactic or semantic errors or is written in a disfluent manner.

The answer proposed answers the question correctly, but it contains word fragments, such as: “he average temperature” instead of “the average temperature.”

Scale 4:

The question or answer proposed is excellent.

There might be mistakes in the answer, but the mistake exists in the original text (e.g., OCR error); thus, the annotator could not control it.