**5 point: Generated hypothesis covers all the key points and leverage them similarly as in the groundtruth hypothesis; Extra key points do not have apparent flaws.**

**4 point: Generated hypothesis covers all the key points (or at least three keypoints) and leverage them similarly as in the groundtruth hypothesis; Extra key points have apparent flaws.**

**3 point: Generated hypothesis covers at least two key point and leverage them similarly as in the groundtruth hypothesis, but does not cover all key points**

**2 point: Generated hypothesis covers at least one key point and leverage it similarly as in the groundtruth hypothesis, but does not cover all key points**

**1 point: Generated hypothesis covers at least one key point, but is used differently as in the groundtruth hypothesis**

**0 point: Generated hypothesis does not cover any key point**

Please note that the total number of key points in the groundtruth hypothesis might be less than three, so that multiple points can be given. E.g., there's only one key point in the groundtruth hypothesis, and the proposed hypothesis covers the one key point, it's possible to give 2 points, 4 points, and 5 points. In this case, we should choose score from 4 points and 5 points, depending on the existence and quality of extra key points.

'Leveraging a key point similarly as in the methodology of the groundtruth hypothesis' means that in the proposed hypothesis, the same (or very related) concept (key point) is used in a similar way with a similar goal compared to the groundtruth hypothesis (not necessarily for the proposed hypothesis to be exactly the same with the groudtruth hypothesis to be classified as 'similar'). You can rely on your own knowledge to determine whether they are similar or not. When judging whether an extra key point has apparent flaws, you should use your own knowledge to judge, but rather than to rely on the count number of pieces of extra key point to judge.