Task	Prompt
Stage 1 Prompt for AQA	Question: {question} \nLable: {label} \n Prediction: {prediction} \nPlease judge whether the prediction is correct. If the prediction matches the label clearly, then the prediction is correct, otherwise wrong. The answer (Yes or No) is: "
Stage 1 Prompt for STF and LTF	Following natural language inference, please judge whether the prediction is in accord with the given labels. **Rules:** \n1. Based on the given labels only. \n2.Note that the tense of the prediction should be disregarded. \n3.If certain label can clearly demonstrate the prediction, then the prediction is correct, otherwise wrong. \n4. If the prediction is wrong, output "No". If the prediction is correct, please output "Yes" along with the corresponding index in the list of labels, for example: "According to the given information, the answer is Yes, and the index is: (1)". \nThe labels are {label}\n The prediction is {prediction} The answer is:
Stage 1 Prompt for STF and LTF	Following natural language inference, please judge whether the prediction is in accord with the given label. Rules: \n1. Based on the given label only. \n2.Note that the tense of the prediction should be disregarded. \n3.If certain content from label can clearly demonstrate the prediction, then the prediction is correct, otherwise wrong. \nThe label is {label}\n The prediction is "{prediction}" The answer (Yes or No) is:

Table 13: Prompt templates for LLM-based evaluation. The STF and LTF adopt the RAE with two-stage evaluation. The stage 1 (step 3 in 5) of RAE evaluates the prediction with gold answers. The stage 2 (step 5 in 5) evaluates the prediction with retrieved web contents.