A THE DESIGN OF EVALUATION PROMPT

The evaluation prompts are adopted for both *qualification exam* and *peer review* modules (detailedly introduced in Sec 3.2). These prompts are fed to the reviewer (or reviewer candidate) LLMs, allowing them to generate ratings or preferences. In our experiments, we have proposed three different prompt settings (pairwise, 5-level pointwise and 100-level pointwise), and then seperately designed the prompt template for each setting, as the following shows. Here we show the design in XSum dataset under each setting, the design of NF-CATS dataset is almost the same.

A.1 Pairwise setting

###Task: Evaluate two summaries of a given passage and determine which one better summarizes the main points of the passage considering accuracy and conciseness. You only need to output 'one' or 'two' directly to indicate which summary summarizes the passage better.

###Passage: { *passage* }

###Summary one: { *summary 1* }

###Summary two: { *summary 2* }

###Output:

A.2 5-Level pointwise setting

###Task: Evaluate the summary of a given passage and determine how it summarizes the main points of the passage considering accuracy and conciseness. Directly output a number between 1 and 5 to indicate the quality score of this summary:
- 1 means the summary is not relevant to the passage,
- 2 means the summary is neither accurate nor concise but it is relevant to the passage,
- 3 means the summary is only a fair summary of the passage considering accuracy and conciseness,
- 4 means the summary is a good summary of the passage but still has room for improvement in accuracy and conciseness,
- 5 means the summary is a perfect summary of the passage considering accuracy and conciseness.

###Passage: { *passage* }

###Summary: { *summary* }

###Score of the summary:

A.3 100-Level pointwise setting

###Task: Evaluate the summary of a given passage and determine how it summarizes the main points of the passage considering accuracy and conciseness. Directly output a number between 0 and 100 to indicate the score of this summary. The higher the score, the more accurate and concise the summary is.

###Passage: { *passage* }

###Summary: { *summary* }

###Score of the summary: