Thank you for agreeing to evaluate Tree-of-Debate, we really appreciate it! Here are the instructions on how the evaluation process works:

Overview:

Our goal is that given as input, two papers and a high-level topic common between them, Tree-of-Debate aims to construct a paragraph-long comparative analysis of what the similarities and differences are between the two papers. This should be robust to whether the papers cite each other or differ in method or task at any level of granularity.

Your task is to identify pairs of papers which you would like to evaluate the comparative analysis for.

The below instructions look long, but are just pretty comprehensive in case you have any questions.

Instructions:

Step #1: Compiling a list of paper *pairs* you would like to review

ToD Final Dataset

In the above Google sheet *(Tab: Dataset Info)*, fill out a row per paper pair. Here is information on each column:

- *Contributor:* Your name
- *Topic*: A phrase-long description of the common topic between the two papers which you would like to ground the analysis on
    - *Example*: using preferences to train language models for better reasoning
- *Paper #i arXiv Link*: This ideally will be the link to the arXiv PDF, but if the paper is not available on arXiv, then the link to the PDF is fine. It should be publicly available/open-access though.
    - *Example:* https://arxiv.org/pdf/2404.02078
- *Paper #i Title*
- *Paper #i Abstract*
- *Paper #i Introduction*
    - If you find this difficult to parse, I can do it, but just let me know
- *Method (0) or Task (1):* What do you expect to see in the output comparative summary? Do you expect that it will be more method-based comparisons because the papers address a similar task, or do you expect to be more high-level discussion of principles/ideas/motivations because they have different tasks but maybe similar methods (or not)?
    - *Method:* More method-based.
    - Task: More high-level, task based.
- *No Cite (0) or Cite (1)*
    - Do the papers cite each other or not?

Try to keep a somewhat uniform distribution between method vs task and cite vs no cite (e.g., 5 pairs per category: M + no cite, M + cite, T + no cite, and T + cite)

Step #2: You evaluate!

ToD Final Dataset

I will provide you with five different summaries for each pair.

For each summary, please evaluate the following:

- **Breadth (score 0-4 per summary):** Does the summary seem comprehensive and complete?
    - **Score 0/4:** No, the summary misses (MULTIPLE) major points.
    - **Score 1/4:** No, the summary misses a (SINGULAR) major point.
    - **Score 2/4:** Somewhat, the summary misses minor points.
    - **Score 3/4:** Yes, the summary covers the major points, but still is not what I would expect.
    - **Score 4/4:** Yes, the summary is fully comprehensive and complete.

- **Contextualization (score 0-4 per summary):** Does the summary explain and/or justify the posed similarities/differences between the papers, as opposed to just mentioning them?
    - **Score 0/4:** No, the summary is simply extractive– just seems to take different subtopics from each paper and doesn't synthesize them– no justification behind similarities and differences.
    - **Score 1/4:** No, the summary attempts at some level of synthesis, but it is not meaningful.
    - **Score 2/4:** Somewhat, the summary attempts at synthesizing at most one point.
    - **Score 3/4:** Yes, the summary contains meaningful synthesis but only for a minority of points.
    - **Score 4/4:** Yes, the summary contains meaningful synthesis across all major points.

- **Factuality (score 0 or 1 per sentence in summary)** Does the sentence maintain fidelity to the papers with respect to the facts?
    - 1 for YES, 0 for NO
        - Say NO if the sentence contains any factual inaccuracy
    - Expected format (for a 5-sentence summary): 0,1,1,0,1