**Welcome to the annotation project for PeerSum**

Please have a careful read of the project introduction and task instructions and finish the tasks in the separate document.

**Introduction of the project**

To enhance the capabilities of multi-document summarization systems we present PeerSum, a novel dataset for automatically generating meta-reviews of scientific papers based on reviews and discussions in the peer-reviewing process in https://openreview.net/. In the reviewing process, all assigned reviewers and public users can give comments to each paper, and then the author of the paper might respond to those comments. There may be a couple of rounds of discussions or rebuttals during the reviewing process. In the end, the meta-reviewer will write a summary of these comments and discussions, to support their final decision on the paper acceptance. Usually, meta-reviewers are supposed to write the meta-review based on summarizing all reviews, discussions, and the paper abstract, but they may sometimes have their own judgements and also draw on some external knowledge for their judgements in meta-reviews.

The objective of the annotation task is to assess whether the **statements/assertions** in meta-reviews are exclusively drawn from the **reviews, discussion and the paper abstract** which are the source documents in PeerSum to generate the content of the corresponding meta-review.

**Instructions for the task**

You will get 6 samples in total. For each sample:

1. Please carefully read the meta-review (i.e., the section of Paper Decision), and source documents including the paper abstract, reviews by different reviewers, and discussions between reviewers and the author (all responses) in the linked OpenReview page.
2. Please highlight **all assertions or statements (which may be a clause, a sentence, or a paragraph)** which draws knowledge solely from source documents with the colour of turquoise. Highlighted texts will not be based on meta-reviewer's judgements or own knowledge in the field, reading the full paper themselves, or any other external knowledge.

**Annotation examples**

We also prepare explanations for unhighlighted texts following each example, but you do not need to write explanations when annotating.

**### Example one ###**

**Source Documents:**

Link to OpenReview: https://openreview.net/forum?id=H1DkN7ZCZ

**Meta-review:**

Authors present a method for representing DNA sequence reads as one-hot encoded vectors, with genomic context (expected original human sequence), read sequence, and CIGAR string (match operation encoding) concatenated as a single input into the framework. Method is developed on 5 lung cancer patients and 4 melanoma patients. Pros: - The approach to feature encoding and network construction for task seems new. – The target task is important and may carry significant benefit for healthcare and disease screening. Cons: - The number of patients involved in the study is exceedingly small. Though many samples were drawn from these patients, pattern discovery may not be generalizable across larger populations. Though the difficulty in acquiring this type of data is noted. – (Significant) Reviewer asked for use of public benchmark dataset, for which authors have declined to use since the benchmark was not targeted toward task of ultra-low VAFs. However, perhaps authors could have sourced genetic data from these recommended public repositories to create synthetic scenarios, which would enable the broader research community to directly compare against the methods presented here. The use of only private datasets is concerning regarding the future impact of this work. – (Significant) The concatenation of the rows is slightly confusing. It is unclear why these were concatenated along the column dimension, rather than being input as multiple channels. This question doesn't seem to be addressed in the paper. Given the pros and cons, the committee recommends this interesting paper for workshop.

**Explanations**

1. "However, perhaps authors could have sourced genetic data from these recommended public repositories to create synthetic scenarios, which would enable the broader research community to directly compare against the methods presented here. The use of only private datasets is concerning regarding the future impact of this work." , this is meta-reviewer's judgement about impact of the paper.
2. "This question doesn't seem to be addressed in the paper." is a meta-reviewer's judgement about the paper.
3. In "Given the pros and cons, the committee recommends this interesting paper for workshop.", there is external knowledge about the workshop information.

**### Example two ###**

**Source Documents:**

Link to OpenReview: https://openreview.net/forum?id=ZeE81SFTsl

**Meta-review:**

Dear authors, I apologize to the authors for insufficient discussion in the discussion period. Thanks for carefully responding to reviewers. Nevertheless, I have read the paper as well, and the situation is clear to me (even without further discussion). I will not summarize what the paper is about, but will instead mention some of the key issues. 1) The proposed idea is simple, and in fact, it has been known to me for a number of years. I did not think it was worth publishing. This on its own is not a reason for rejection, but I wanted to mention this anyway to convey the idea that I consider this work very incremental. 2) The idea is not supported by any convergence theory. Hence, it remains a heuristic, which the authors admit. In such a case, the paper should be judged by its practical performance, novelty and efficacy of ideas, and the strength of the empirical results, rather than on the theory. However, these parts of the paper remain lacking compared to the standard one would expect from an ICLR paper. 3) Several elements of the ideas behind this work existed in the literature already (e.g., adaptive quantization, time-varying quantization, ...). Reviewers have noticed this. 4) The authors compare to fixed / non-adaptive quantization strategies which have already been surpassed in subsequent work. Indeed, QSGD was developed 4 years ago. The quantizers of Horvath et al in the natural compression/natural dithering family have exponentially better variance for any given number of levels. This baseline, which does not use any adaptivity, should be better, I believe, to what the author propose. If not, a comparison is needed. 5) FedAvg is not the theoretical nor practical SOTA method for the problem the authors are solving. Faster and more communication efficient methods exist. For example, method based on error feedback (e.g., the works of Stich, Koloskova and others), MARINA method (Gorbunov et al), SCAFFOLD (Karimireddy et al) and so on. All can be combined with quantization. 6) The reviewer who assigned this paper score 8 was least confident. I did not find any comments in the review of this reviewer that would sufficiently justify the high score. The review was brief and not very informative to me as the AC. All other reviewers were inclined to reject the paper. 7) There are issues in the mathematics – although the mathematics is simple and not the key of the paper. This needs to be thoroughly revised. Some answers were given in author response. 8) Why should expected variance be a good measure? Did you try to break this measure? That is, did you try to construct problems for which this measure would work worse than the worst case variance? Because of the above, and additional reasons mentioned in the reviewers, I have no other option but to reject the paper. Area Chair

**Explanations**

1. “Dear authors, I apologize to the authors for insufficient discussion in the discussion period. Thanks for carefully responding to reviewers.”, this is coordination words from the meta-reviewer.
2. “Nevertheless, I have read the paper as well, and the situation is clear to me (even without further discussion).”, this is based on meta-reviewer’s own reading of the full paper.
3. “I will not summarize what the paper is about, but will instead mention some of the key issues.”, this is coordination words from the meta-reviewer.
4. “1) The proposed idea is simple, and in fact, it has been known to me for a number of years. I did not think it was worth publishing. This on its own is not a reason for rejection, but I wanted to mention this anyway to convey the idea that I consider this work very incremental.”, this is based on the meta-reviewer’s own experience.
5. “In such a case, the paper should be judged by its practical performance, novelty and efficacy of ideas, and the strength of the empirical results, rather than on the theory. However, these parts of the paper remain lacking compared to the standard one would expect from an ICLR paper.”, this is the meta-reviewer’s experience about the standard of ICLR.
6. “which have already been surpassed in subsequent work. Indeed, QSGD was developed 4 years ago. The quantizers of Horvath et al in the natural compression/natural dithering family have exponentially better variance for any given number of levels. This baseline, which does not use any adaptivity, should be better, I believe, to what the author propose. If not, a comparison is needed.”, this is based on the meta-reviewer’s experience in the field.
7. “5) FedAvg is not the theoretical nor practical SOTA method for the problem the authors are solving. Faster and more communication efficient methods exist. For example, method based on error feedback (e.g., the works of Stich, Koloskova and others), MARINA method (Gorbunov et al), SCAFFOLD (Karimireddy et al) and so on. All can be combined with quantization.”, this is based on the meta-reviewer’s experience in the field.
8. “I did not find any comments in the review of this reviewer that would sufficiently justify the high score. The review was brief and not very informative to me as the AC.”, this is meta-reviewer’s judgement on the review.
9. “7) There are issues in the mathematics – although the mathematics is simple and not the key of the paper. This needs to be thoroughly revised.”, this is based on meta-reviewer’s reading of the full paper.
10. “8) Why should expected variance be a good measure? Did you try to break this measure? That is, did you try to construct problems for which this measure would work worse than the worst case variance? Because of the above, and”, this is based on the meta-reviewer’s own knowledge in the field.
11. “I have no other option but to reject the paper. Area Chair”, this is the meta-reviewer’s judgement on the paper.