Dear Reviewers,

We thank you for your valuable feedback to help us improve the quality of our research paper. To address the initial questions about why engagement is needed in a pedagogical tool, and why we are interested in predicting the student engagement, we included the main reason affecting the quality of the peer groups in the Introduction section. We also addressed limitations and biases in the methodology. The discussion points are highlighted below.

Student engagement is a crucial part of the peer learning process because peer education is most effective when every student in the peer group participates and engages with each other bringing to light diverse points of view of solving the same question. This facilitates learning, critical thinking, and healthy debates to find the right answer in the group discussion. By predicting student engagement, we can identify which cyber peer-led team learning (cPLTL) groups are performing consistently well, which groups need help and which peer leaders need additional support from the educators. Currently, in the cPLTL model, during the online recitation, educators are absent. This means that they do not know anything about the performance of the peer group. So, the machine learning model serves as a prediction tool to help provide feedback to the educators about their peer groups. These insights about student engagement are necessary to assess the quality of the cPLTL progression over the course of a semester.

Yes, the same prompts were used in both ChatGPT and Bard to determine whether these generative AI tools could perform well when compared to the machine learning model.

To validate the results from generative AI tools, we compared the sentiment from traditional machine learning as well as the scores from the two independent subject matter experts. They matched up. Also, the scores from the two human experts were averaged to ensure consistency in the ratings. In future, we will consider using a panel of 3-4 human experts who will validate the results from ChatGPT and Bard. This will also provide deeper insights into how effective the machine learning tools are compared to the results expected from the human experts.

For the preprocessing steps in Python, we first removed all personal identification such as participant's names and personal information that may have been discussed in the transcript. We also used Python's natural language tool kit's (NLTL) built-in functions such as lemmatization, stop words, removing stem words and punctuation. This improved the sentiment scores as it removed the noise from the transcripts.

The limitations and biases of using ChatGPT and Bard were that both gave biased answers for long structured sentences. It appears they performed better on short sentences in the transcript. Another limitation is that the training data or version of ChatGPT was up to September 2021 in the older version. We do not know if this had any impact on the results. We will run experiments on newer generative AI tools. So, generative AI needs strong validation from human experts. Also, since we are using domain specific content from Chemistry, ChatGPT and Bard were unable to interpret nuanced content from specific Chemistry content like equations. However, they were able to gauge participant sentiment correctly. Since sentiment was the main idea in the paper, we were satisfied with the methodology and results. In future, we will consider alternative generative AI tools to predict sentiment.