Dear Reviewers,

We want to thank the reviewers for their thoughtful feedback and comments which helped us readjust and validate our findings and results. We would like to address the comments by the reviewers individually and have put into the letter the specific comment we are addressing in the following section:

Reviewer A

Comment: *An interesting future work is to see the generalizability of training a classifier compared to GPT-based models and ways to improve the GPT-based method performance. It seems the classifier needs to be trained on each subject, despite the authors mentioning "no more data collection is needed". The classifier performance also leaves plenty of room for further improve the performance on this binary classification task.*

Yes our future work will focus on the two thrusts of improving our classifier's performance as well that of the GPT-based method. As improvement in both will lead to a better performance in our combined model. In regards to the statement we claim about "no more data collection is needed" this pertains to our current methodology in which we separate our training by the subjects. The idea is that given this current model if educators wanted to provide further assistance in regards to this specific subject they can use the model to prescreen their hints. However, in the future the idea is that the model will be trained on all questions and hints and not subjected to the boundary of subjects. In this scope, it will lead to a a model that works with any type of subject and can be used as a true pre-screener for all types of hints as long as it's a subject in which it has been trained on, is the hope.

Reviewer B

Comment: *1.) Why do you choose to compare only between two assistance and not among several of them?*
2.) More discussions on why GPT-3.5 and 4 performed worse than your method would be important to make sure that your neural network is just not overfitting on the data.
3.) A case study of a real example involving a question and the assistances chosen by the baseline models and your models could be included for better demonstration.
4.) There is a typo in section 4.2 - GPT 3.5 and 4.5 is mentioned. I believe that it should be 4 and not 4.5.

We will address the comments by the index as the reviewer has asked them:

1.) The reason we choose to compare only two assistance for the following reasons: The goal is to decide which hint is better, rather than predicting the actual reattempt correctness rate (which we had tried out initially but was a harder task). A voting consensus mechanism can be adapted to pick out the

best hint within the set of 4-6 hints typically available for each question, and Pairwise comparison is an easier and more interpretable task.

2.) GPT3.5 and 4 performance is mostly dependant on the prompting is what we discovered through experimentation. In the future we would like to experiment with additional prompts and tuning, however from the prompts that we have tried so far the best result was what we showed in our results and the associated prompt. The model is not overfitting due to the fact that we split up the training and test datasets by the question pairs. Therefore in the test dataset the model is only exposed to questions and hints that it has never seen before.

3.) Due to limits in the page requirements we decided to not included this but will do so in our presentation as well as the poster.

4.) Thanks for pointing that out - in our paper we actually only used GPT 3.5 and 4. This was due to the fact that those were the only ones available at the time. In the future we plan to test out the performance with GPT4.5 as well.

Yours sincerely,

Ted, Harshith, Robin, Amos and Tom

Comments from Reviewer 1:

Comment

Our response

Comment

Our response

Comments from Reviewer 2:

Comment

Our response

Comment

Our response