# COVER LETTER

We sincerely thank the Editors and Reviewers for taking the time to review our manuscript. Your insightful comments and feedback have been highly valuable, and we have improved the manuscript accordingly.

## ***Reviewer 1:***

**Summary** *Binary assessment questions are limited in their ability to evaluate students, and therefore, there is a need for constructed response questions. However, these more open-ended questions require human evaluation, which can be time-consuming. This paper introduces and evaluates a hybrid neural network (HNN) model for open-ended automated assessment of a science question about applying chemistry ideas in the real world. They compare their model (a multi-perspective/perceptive HNN) to a Naive Bayes model, a logistic regression model, a BERT transformer, and something called AACR. They find their approach was "substantially more accurate than the other algorithms".*

**Authors' Reply.** We appreciate that the reviewer understood the paper and correctly summarized it.

**Cons 1:** I think the authors spent a long time building up their motivation to use a hybrid neural network for automated assessment, but unfortunately, I found the description and definition of the HNN quite lacking. For example, what exactly are the inputs to the model? How long were the student responses? And what exactly were the output labels the model is trained to predict?

**Authors' Reply.** We have clearly defined and described the HNN in the revised version on pp. 5 and also provide an architecture diagram as Fig 2. Furthermore, there are specific answers to the raised questions. For example, the input to the models are word tokens, and on average, student responses are 150 characters long. The model finally outputs an automatic score based on scoring perspectives mentioned in Table 2 in the revised manuscript.

**Cons 2:** The authors mention a hybrid neural network combining symbolic approaches with a neural network, but I'm not sure where the symbolic approach is in their definition. They describe a model using BERT word embeddings, a Bi-LSTM layer, and an attention mechanism.

**Authors' Reply.** We have used a symbolic approach as knowledge representation using word embedding in a hybrid architecture. It can be seen in the revised version's HNN definition on PP. 5 and Fig 2.

**Recommendation:** The accuracy scores reported in Table 2 seem quite high already (even for logistic regression). It might be worth selecting question items that are more difficult for older ML methods; otherwise, there isn't much room for improvement.

**Authors' Reply.** We agree with your suggestion and will accommodate more complex question items in the future. However, even in the current manuscript, we have achieved 8% higher accuracy than other approaches, which is a significant improvement for automatic scoring.

**Comment 1:** In the abstract, "their evaluation" can be read as ambiguous. I would explicitly say that "human evaluation" is time consuming (unless you mean automatic evaluation?)

**Authors' Reply.** We have revised the abstract to make it clear to understand in the revised version.

**Comment 2:** You mention that prior automatic evaluation uses "outdated models" and that these are insufficient for "current deep learning standards". I'm not sure what this statement is really adding. Are the outdated models performing poorly on current datasets? In your paper, the new method only performs a little bit better compared to logistic regression.

**Authors' Reply.** We referred to the traditional machine-learning approaches, such as Naive Bayes and Logistic Regression, as underperforming for complex science questions. In this study, we have achieved, on average, 8% and 10% higher accuracy than Naive Bayes and Logistic Regression, respectively. Furthermore, We have shown that traditional approaches don't accommodate multi-perspective scoring, which is important for explainability and provides more insight to the teacher for automatic scoring.

**Comment 3:** In the abstract, you say "aiming to enhance accuracy". You can state whether or not your method worked in the abstract. Perhaps conclude the abstract with your main finding.

**Authors' Reply.** We appreciate your suggestion and have revised the abstract by stating that the approach worked and providing quantitative analysis for comparative study against other machine learning approaches.

**Comment 4:** In the introduction, you say "a prospective solution... using machine learning algorithms". Prospective sounds like it hasn't been tried yet. Older ML methods are still ML.

**Authors' Reply.** We have revised the introduction to understand the narrative better and avoid phrases like "prospective solutions".

**Comment 5:** Typo: "ANNs use artificial neural [networks] to represent".

**Authors' Reply.** Thank you for highlighting the typos; we have proofread the manuscript and fix typographical errors.

**Comment 6:** In the introduction and elsewhere in the paper, you connect ANNs to biological neural networks in the brain. I would urge caution on making this connection. ANNs were inspired by the brain, but that's largely where the connection ends. They don't "mimic" them.

**Authors' Reply.** We have clarified the ANN connection with biological neural networks in the introduction and revised the narrative for better understanding in the revised manuscript.

**Comment 7:** Potential typo: "trained a SciEdBERT to achieve students' written responses". Correct usage of "achieve"?

**Authors' Reply.** Thank you for highlighting the typos; we have fixed them in the revised version.

**Comment 8:** "We innovatively developed and applied a hybrid neural network (HNN) to automatically score student responses using an analytic scoring rubric to address the concerns

of automatic scoring using the analytic scoring rubrics" —¿ this sentence almost sounds like you're saying the same thing twice. Do you mean automatic scoring using analytic rubric WITH HNN versus WITHOUT HNN? I.e. one method only uses the analytic rubric?

**Authors' Reply.** In this study, we have proposed an HNN approach for multi-perspective automatic scoring and compared the accuracy with other approaches that don't use analytic rubrics and mainly focus on binary classification. We also have revised the narrative in the introduction for better understanding.

**Comment 9:** In Section 2, at the end of the first paragraph, I would add a comma after "level", or rearrange the sentence slightly: "However, [this study noted... ], [it could not deliver] [when groups of items]" (just a suggestion).

**Authors' Reply.** Thank you for the suggestion; we have properly accommodated your suggestion in our writing.

**Comment 10:** Towards the end of Section 2, you have a description of ANNs (beginning with "So named for..."). I'm not sure whether this is needed.

**Authors' Reply.** We have excluded the ANN definition from the end of section 2 in the revised version and properly used ANN terminology in the subsequent sections.

**Comment 11:** At the beginning of Section 3, you describe CNNs and RNNs. Combining both does not yield an HNN (symbolic + ANN). It's just another ANN.

**Authors' Reply.** We have addressed this confusion in the revised version and properly defined HNN with the combination of ANN with symbolic or knowledge representation approaches.

**Comment 12:** Section 3, the sentence beginning "HNN has been used in the" is quite long, and I found it confusing.

**Authors' Reply.** We have broken down that sentence into two parts for better understanding in the revised version.

**Comment 13:** You change between multi-perspective and multi-perceptive.

**Authors' Reply.** It was typographical error, the original term is multi-perspective, and we have fixed this error in the revised version by only using "multi-perspective"

**Comment 14:** I'm not familiar with the term "perspective" for the different scoring labels. This confused me quite a lot because I when you refer to multi-perspective neural networks, I thought this was a special type of NN, rather than referring to the output labels.

**Authors' Reply.** The scoring perspective is defined in Table 2, which is different than multi-perspective neural networks. We have used it as an output of HNN based on the rubric given in Table 2 rather than having it as a special type of NN.

**Comment 15:** In Section 4, how large was the training dataset? You say it towards the end of the section (80% training is 1K human responses), but I think you should say it earlier when you mention the "sizable corpus".

**Authors' Reply.** We have mentioned the size of human responses in Section 4 (pp. 4) at the

beginning, along with the participants' descriptions in the revised version.

**Comment 16:** What is AACR? It wasn't defined.

**Authors' Reply.** AACR was one of the neural networks used for automatic scoring; we have changed AACR to ANN for simplicity in the revised version.

**Comment 17:** I'm not sure what the "scoring aspect" means in Table 2, it wasn't explained.

**Authors' Reply.** Table 2 of the revised version defines the rubric and mentions associating the scoring aspect with perspective; for example, scoring aspect one means the student response contains SEP and DCI components.

***Reviewer 2:***

**Summary** *The paper proposes a scoring algorithm to assess students' science knowledge. The paper has a novel approach and technical contribution.*

**Authors' Reply.** We appreciate the time and effort the reviewer put into reviewing the manuscript and appreciate the constructive feedback.

**Comment 1:** However, the writing style of the paper is hard to follow. For example, the Introduction is written more like a Related Work section. It would have been easier for readers to understand the paper if the research questions were explicitly written at the beginning of the Introduction. It is difficult to understand what authors have done by reading the paper. Writing can be improved.

**Authors' Reply.** We have revised the format of the paper and properly mentioned research questions in the introduction based on the feedback. Furthermore, the revised version has also been thoroughly proofread to make it understandable to the readers. We hope the revised version is easy to follow and understand.