**Response to Reviewer #1:**

Q1:The proposal of hierarchical scorer (sentence→paragraph→essay) and trait-specific features could seem reasonable. However, the paper didn't present any motivating examples to backup the design of the architecture

A1:The reason why we use the hierarchical architecture (sentence→paragraph→essay) is that for writing, it includes sentence level, paragraph level and essay level. Each level is the premise of the next level. For a paragraph, it is necessary to write a sentence well first, and the premise of a good article is also to write effective sentences first. Therefore, we design a three-level architecture from low to high. For trait-specific, our design motivation is that the criteria for judging the quality of a certain trait of an essay are different, so we expect the addition of features to show this difference.

Q2:Beautiful Sentence Scorer seems to be an important component. The paper didn't provide enough details on the design and implementation.

A2:The design details are shown in lines 412-437 of the article. We divide the dataset according to the ratio of 8:2 and only use XlNet to encode sentences to train a beautiful sentence grader through the dense layer. The final F1 value can reach 0.91. The specific implementation is that the results of the dense layer are normalized as features and directly spliced with language expression as features.

Q3:From the ablation study, the removal of Logic led to a highest drop in performance. This reflect some kinds of weights (or importance factors) among the 4 traits. Are these weights consistent with the essay assessment guideline in the manual essay scoring? Can any manual guidelines be used as priors for AES models?

A3:In the criterion for the essay of college entrance examination, there is no standard to rank the importance of these four traits, so we consider the same weight in our experiment, without considering a priori. If we can have a priori guide, the effect may be better.

**Response to Reviewer #2:**

Q1:A majority of references is repeated in almost the same way at the beginning of the intro as and in Section 2.1. Either it could be explained more in 2.1. or there might be more space, e.g. to explain more above the above mentioned issues.

A1:Due to the limited space, we have not given detailed explanation. If there is a longer space, we will consider more explanations.

Q2:You could also say a bit more about the dataset: How many prompts? Which topics? Who wrote the essays?

A2:Our dataset does not contain prompts, which makes the task more challenging. At the same time, our dataset includes articles on various topics, which are argumentative essays written by high school students screened from the Internet.

Q3:Is the IAA kappa mentioned in 3.3 also QWK as used later in the experiments?

A3:They are different. The QWK evaluated in the follow-up experiment is the twice weighted kappa value, which is widely used in the field of automatic essay scoring.

Q4:line 468 talks about the number of paragraphs per paragraph, there must be some typo here.

A4:I'm sorry I didn't express it carefully. This sentence should be “The maximum numbers of sentences and paragraphs in any essay(i.e. n, m and np) are set to 50 and 20, and the maximum number sentences in any paragraph is set to 20.

Q5:Section 5.3 talks about baselines but then lists only 4, probably as two models are used in two conditions each. That could be formulated in a clearer way.

A5:Thank you very much for your comments. We will express them more clearly in the subsequent amendments.

Q6:The paper talks about significant results. Please provide the details about significance tests, level of significance etc.

**Response to Reviewer #3:**

Q1:Although the proposed HMTS achieves outstanding performance on the collected dataset, the techniques used in it (i.e., multi-task learning, hierarchical architecture for article modeling, and attention mechanisms) are not new.

A1:Our innovation is not new techniques but the design of architecture. Essay scoring is a partial linguistic research, and most of the latest research also lies in architectural innovation. Our innovation is that we are the first to apply multi-task learning to Chinese essays without prompts and make each trait learn from each other through sequence attention. The trait specific-feature also strengthens the representation of each trait.

Q2:The annotated dataset contains 1220 essays, which is relatively small. In addition, the number of double-annotated essays to compute the inter-annotator agreement is unclear, which is important in determining the annotation quality.

A2:Most existing datasets are in English and prompt specific, while our dataset is in Chinese and does not contain prompts. At present, this research is still in development, and there is no large dataset.

Q3:The four scoring aspects (i.e., organization, topic, logic, and language) used in the studies are determined arbitrarily. It would be better to discuss the motivation of the design (e.g., why the authors do not use other aspects such as content consistency).

A3:We selected the four most important aspects according to the essay evaluation criteria of the national college entrance examination, which is the decisive examination for Chinese students to enter universities.

Q4:The 6th paragraph of the Introduction (starting from line 77) discusses the differences between Chinese and English essay scoring, which is somewhat counterintuitive. Generally, sentence- and paragraph-level expressions, as well as smooth expressions, are also important for an English essay. Therefore, more concrete examples or citations of existing studies are necessary to make the claim clearer.

A4：Maybe there is a problem with my statement. What I want to discuss is that Chinese expression pays less attention to words than English, and pays more attention to the expression of sentences and paragraphs. Therefore, the evaluation criteria should be different, not completely different. For example, the use of long and difficult words or high-level words will add luster to English essays, while there are no long and difficult words in Chinese. Simple words can also make beautiful sentences, such as "A bed, I see a silver light, I wonder if It's frost around.", which was written by the famous Chinese poet Li Bai. Simple words alone form beautiful sentences that have been handed down through the ages.

Q5:In Section 3.3, how many essays are annotated by both annotators?

A5:We have two annotators who both annotate 1220 articles.

Q6:Do the authors plan to release the annotated dataset?

A6:If our paper can be accepted, we will publish the annotated dataset.