***Response to Reviewers’ Comments***

**Ms. Ref. No.:** DIB-D-23-01547

**Title:** An Applied Statistics dataset for human vs AI-generated answer classification

# Summary

We are grateful to the anonymous reviewers for their detailed observations and constructive feedback. We have revised the manuscript considering all the issues. Responses to the specific review comments (shown in italic) and other changes to the manuscript are given below.

***Issue*:** *One thing that I noticed which was missing is the consent form. This should be done when the user registers so as to allow the collector to get the user's consent.*

**Response**: We took consent in paper format from the participating students before collecting data from them. We have also revised the software to take consent from users when they upload the answer.

***Issue****: How did you make sure that the students did solve the question on their own and did not use AI tools like ChatGPT to generate the answers? As you stated in the motivation, current detection tools do not work well, so how did you make sure that the students did not use AI to generate the answers? Quality control strategies on collecting answers from volunteer students.*

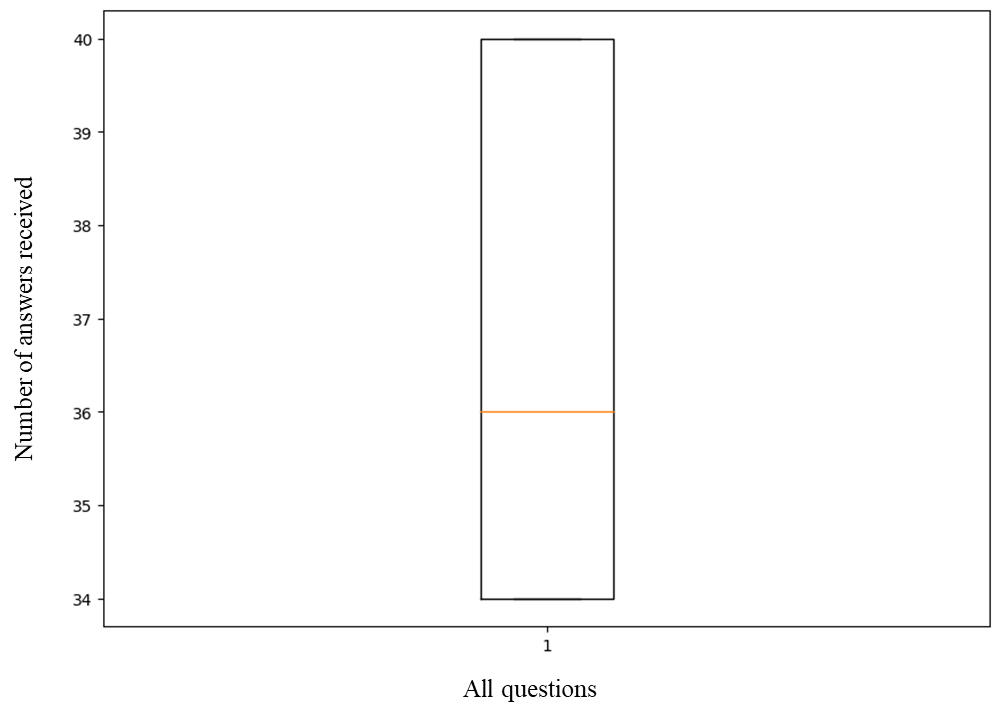
**Response**: Each student volunteered for this without financial or academic gain. We also trained several classifiers using the 80-20 train-test split of the dataset and achieved good performance (page 3, Table 1 in the revised manuscript) from most of them which further indicates that the students performed their tasks properly.

***Issue*** *: I saw at the end that you used Turn-it-in for plagiarism check, but I was not sure what the context of "plagiarism" here means?*

**Response**: We used this because we need to ensure that the responses provided by the students differ from each other and also the AI-generated responses are distinct.

***Issue*** *:* *Given that the each student randomly answered 50 questions, how did you guarantee that each question had the same amount of answers at the end? You could easily end up with unbalanced data-set, where some questions would have more answers than others? I think it would be good if the tool can balance the random selection so that you get similar or equal answers at the end.*

**Response**: We have introduced new features in our software to balance the data by making sure each question is randomly assigned to a similar number of students (as described in revised manuscript page 3, Eq. (1)). We have also tried to balance our dataset by collecting new answers from students which received fewer answers before. In our dataset, number of answers received for each question ranges from 34 to 40. Some variations have occurred because some answers received fewer and duplicate answers from the student. Also, some inappropriate answers were removed. The final summary of the distribution is given in following boxplot. It is a five-number summary of the data: Min: 34.000 Q1: 34.000 Median: 36.000 Q3: 40.000 Max: 40.000.



***Issue****: Analysis on the differences between student-authored answers vs. ChatGPT-generated answers in terms of the number of unique words, the average length of answers, the diversity for answers for the same questions, etc.*

**Response**: The analysis of the differences between human and AI answers in your dataset yielded the following results:

Table 1. Diversity of the dataset

|  |  |  |
| --- | --- | --- |
| Generated answers | Number of unique words | Average length of answers  (in words) |
| AI | 10318 | 151.2 |
| Human | 11801 | 77.3 |

These findings suggest that, on average, AI responses are longer than those from humans. Furthermore, compared to AI responses, human responses show a wider range of unique terms, indicating a possibly higher diversity in vocabulary usage.

***Issue:*** *Show Benchmark results of a baseline detection tool trained on the curated dataset.*

**Response**: The benchmark results are shown in Table 2 of the revised manuscript.

***Issue:*** *In the manuscript, the authors mentioned on several occasions "selected by a domain expert”, I assume that this means a single person choosing the questions. I believe that this is rather small, and it would have been better to get more experts involved in collecting the questions. Additionally, some details are missing, things like: what were the expert qualifications and what they were suited for this task? Based on what did they select the questions? What were the selection criteria?*

**Response:** “A domain expert” was a typing mistake. Actually, two teachers created questions in this study. The experts have Ph.D. degrees and have been teaching this subject at undergraduate and postgraduate levels for more than five years. The teacher selected the questions to cover a wide range of topics from applied statistics.

***Issue*:** *Finally, a suggestion from my side is maybe to add some examples on how such data can be useful by citing some papers that either have used similar data sets within the*

*same contexts. There were some recent papers that came out that had used something similar for both using the data and training AI detection tools:*

*- Ibrahim, H., Liu, F., Asim, R. et al. Perception, performance, and detectability of conversational artificial intelligence across 32 university courses. Sci Rep 13, 12187 (2023).*

*https://doi.org/10.1038/s41598-023-38964-3*

*- Vasilatos, Christoforos, Manaar Alam, Talal Rahwan, Yasir Zaki, and Michail Maniatakos. "HowkGPT: Investigating the Detection of ChatGPT-generated University Student*

*Homework through Context-Aware Perplexity Analysis." arXiv preprint arXiv:2305.18226 (2023).***Response:** We have added those citations in the revised manuscript.

***Issue*:** *Have you obtained an Institutional Review Boards (IRB). This is to ensure that you comply with applicable regulations, meet commonly accepted ethical standards, follow institutional policies, and adequately protect the research participants.*

**Response:** We collected consent from the students and there is no identity-revealing information in our dataset. We don’t have any IRB board in our institute.

***Issue*:** *Also, there were a couple of small issues and typos here and there, please make sure to go over the text and fix them. Here are two of those that I noted:*

*- an equal no --> an equal number*

*- AI detector tools --> AI detection tools***Response:** We have thoroughly checked the manuscript and corrected all the typographical errors that we found.