Point-To-Point Responses

We would like to thank the reviewers **R1**, **R2**, **R3** for their thorough review of our submission Comparing Artificial Intelligence Models with Human Experts for Assessing CVSS from Natural Language Descriptions. Please find our point-by-point response below. For ease of reading, we have included the original reviewer comments under the blue header and our responses within the black 'Response' paragraph header and highlighted the changes in the revised manuscript in boxed text with the original text starting with a minus sign and the revised text with a plus sign.

R1.Q1 First, I believe this paper is both too long and in too many places topo deep in the weeds. The paper gives the impression of a dissertation that was (not enough) cut down to an article [which I assume it is]. I think a lot more cutting could be done on the technical side.

Response It's the other way around: this AI-focused journal submission extends a shorter conference paper (Zhang, Zijing, et al. "Assessing Vulnerability from Its Description." International Conference on Ubiquitous Security. Singapore: Springer Nature Singapore, 2022.), instead of cutting down from a dissertation, which will contain three other papers on CVE description labeling, CVE classification, and a conclusive work on alternative vulnerability database.

R1.Q2 Conversely, there is too little information on the comparison with human judgments. The paper states that " A comparative study between the deep learning models and the human domain experts shows that the deep learning models outperform the human experts in the CVSS estimation task by a large margin." Unfortunately, there is way too little in the paper to justify that conclusion. What was compared? What were the input data? What was ground zero?

Response We've compared with the results from our former conference paper with the paper "An expert-based investigation of the Common Vulnerability Scoring System". Namely, Table 5, Figure 14, 15, 16, 17, 18, and Section 6.3 provides evidence to conclude that AI model can outperform/outscale human significantly. The human-side input data are 384 expert surveys (Figure 7) to evaluate 3000 vulnerabilities in the paper aforementioned by Holm et al, as referred by Section 3.3. The input data for the AI model are 118k NVD entries as training data and 51,000 entries as test data as detailed in Section 5.1. What are ground zeros? Is this reviewer trying to say ground truths? If so, they are the CVSS ratings by the NVD contractors.

R1.Q3 Even a non-expert like I should be able to answer those questions, but, try as I might from reading and re-reading Section 6.3, there are two many questions and entirely too little reflection on the methodology and limitations of the Holm & Afridi paper. In fact, that paper is held up as THE comparison, but going to it, we find a whole lot of good reasons why the errors are higher: First, 304 experts were used, each of whom did only 10 analyses. The number of experts alone increases variability when compared to one (1) algorithm. How about comparing 304 algorithms, all trained on different subsets of data, to 304 experts? Second, the human experts each only rated a few examples, and only 3 vulnerabilities were rated by all; the algorithm had a much larger sample size.Basic psychometrics tell us that longer tests (i.e., tests with more items) have greater reliability and validity than shorter ones. Samples based on fewer examples will thus have greater variability and error. So, what to do? I believe this paper should de-emphasize the human comparison (a lot!) or should give much more space and thoughtful elaboration to the comparison with the human experts.

Response This reviewer probably mean 384 experts, not 304 experts. Design flaws of the CVSS metric, human expert bias, and the small sample size of the expert survey contribute to the variance

in the expert survey. Our AI model aims to mitigate human bias by providing a more conclusive, generalized, and reliable CVSS estimation than the human counterparts through the use of a large dataset, a unified model, and a consistent training process whereas the human experts could have different interpretations of the CVSS metric, different domains of expertise, and different levels of experience. Therefore, training 384 biased models would defeat this research purpose by comparing biased models with biased human experts. Rather, we are trying to do our best effort to compare a generalized, unbiased model with the statistical denoised/unbiased generalization of human experts for estimating cybersecurity vulnerability.

Title Change The title has been changed to reflect the de-emphasis on human comparison.

> -Comparing Artificial Intelligence Models with Human Experts for Assessing CVSS from Natural Language Descriptions
> +AI-Enabled Automated Common Vulnerability Scoring from Common Vulnerabilities and Exposures Descriptions

New Problem Statement Section We have added a new section to the paper to emphasize the problem statement and the research objectives so that it is clear that the focus is on AI models. This new section also explicitly states the limitations of the Holm & Afridi paper. Please see the revised manuscript for more details.

Reflection on Human Experts We have added a new paragraph in the Related Works section to the paper to reflect on the methodology and limitations of the Holm & Afridi paper.

> +Design flaws in the CVSS formula, as mentioned in the Background section, the small sample size of the survey, the subjectivity of human judgments, and the different opinions from experts from different backgrounds jointly contribute to the variance in the human expert CVSS estimation.
> +Nonetheless, Holm et al. had been the only publicly accessible work to represent the performance of human experts in the CVSS estimation task.
> +Therefore we use their work as a reference to compare the performance of the AI models with human experts.

A New Introduction Section with Less Technicality and a New Big Picture Figure We have added a new introduction section to the paper to provide a high-level overview of the paper and the research objectives.

Reviewer 2 Comment In this paper, the authors compare the performance of models with human experts, motivated by the fact that AI is now used as a substitute for humans to find security vulnerabilities. One of the presented models shows superior performance for assessment tasks for different types of vulnerabilities. The authors also investigate the effectiveness of models in predicting the components and severity of vulnerabilities. The results underline the potential of agents with AI to support cybersecurity experts. The presented work deals with an important topic, is interesting, relevant, fits well in the chosen journal and is also well written. For all of the above reasons, this reviewer is positive and recommends acceptance and makes only a few minor suggestions for improvement to further enhance the benefits for the interested reader. The content of the paper is good and the experiments conducted are conclusive.

R2.Q1 The paper is written in a bumpy English, it is recommended to have the manuscript proofread by a natural English native.

R2.Q2 The figures are very small, Figures 14 to 20 are hard to read - this reviewer printed the paper in B/W on paper - the production department must decide what can be improved here.

Response We changed the size of the mentioned figures. Due to the page limits, we removed some of less important figures and tables to make the paper more concise.

R2.Q3 The doi is missing in all references - but this is essential to get quickly to the respective paper - URLs, on the other hand, should be avoided if possible, e.g. 27 does not work either - and also has no description - you click into the unknown and end up on a broken link

Response We have added the DOIs, if present in crossref.org, to all references in the revised manuscript. After double-checking the URLs, they work well except for the one on effect size (Power analysis, statistical significance, and effect size.) The link functions properly if one clicks "advanced" and then "continue to the website" on the warning page in Chrome. This is due to the website's security certificate as the University of Michigan hosts the webpage.

R2.Q4 At least one paragraph "future work" should be inserted in the conclusion or discussion and describe possible further work, this would be helpful, because there are still some open questions here, for example, one could achieve enormous added value by means of explainable AI methods and answer many more questions, here one can stimulate further work and also give the reader an outlook and also insert a reference, for example, there is a completely new paper that presents some such methods and thus offers a good starting point for the interested reader, see: Retzlaff, C.O. et al. (2024). Post-Hoc vs Ante-Hoc Explanations: XAI Design Guidelines for Data Scientists. Cognitive Systems Research, 86, (8), 101243, doi:10.1016/j.cogsys.2024.101243.

Response We added a Future Works section to the paper to describe possible further work.

> +Future Works
> +Larger expert samples could help to reduce the variance in human estimation and therefore provide a more accurate comparison between the AI model and the human experts.
> +The AI models also lacks the ability to explain the reasoning behind the CVSS score prediction.
> +Provision of the reasoning behind the prediction would be necessary to build trust in the AI model.
> +It is also worth exploring the performance of the AI model on the upcoming CVSS 4.0 scoring system and how effective the AI models in help improving the vulnerability processing pipeline.

We've added the resource consumption for SVR and USE to compare the shallow model and the deep learning model in the result section revised manuscript.

> +Resource Consumption (Video peak usage memory in GBs; time in minutes; storage in GBs) for USE and SVM: the USE-V2 model trains & inferences faster than the USE-Large model but consumes more memory and storage.
> +The SVM costs no VRAM to run and only requires 277 MB of main memory at peak

usage.
Table of resource consumption shows the resource consumption to train and test the USE and SVM models on a P100 instance on Kaggle with 30699 test inferences and 92097 training samples one epoch on a CVSS prediction task with a batch size of 9. +The USE-V2 model consumes 3.06 GB of video memory and 1.04 GB of storage, which is twice as much as the USE-large variant, which peaks at 3.66 VRAM consumption.
+However, the USE-V2 model trains and inferences 6x faster than the USE-large model.
+The GPT-3 model runs remotely on the OpenAI API therefore there is no available data on its resource consumption.
+The SVR model with the TF-IDF vectorization consumes 225 minutes to train and requires no video memory and runs on the CPU. Even though the SVR has the best performance (0.74 CVSS base score) in terms of the mean absolute error (MAE), the USE model is more efficient in terms of time required to train and inferences.

R3.Q1 The abstract does not currently have sufficient details of what the paper is about. I would like the authors to include more details on the models used, the dataset, and key results (some of the results are provided by authors but I suggest elaborating and giving a good bird-eye view). The abstract should also show some emphasis on the comparison metrics used and the significance of outperforming human experts.

Response We have revised the abstract to include more details on the models used, the dataset, and key results.

-This paper proposes employing artificial intelligence models as substitutes for humans or as aides to human experts in estimating vulnerabilities.
-We compare the performance of two such models with human experts in estimating vulnerability severity scores.
-One of our models demonstrates superior performance, with results that significantly outperform human experts in assessment tasks for various types of vulnerabilities.
-Additionally, we examine the efficacy of our models in predicting the components and severity level of vulnerabilities.
-The findings highlight the potential of artificial intelligence agents to assist cybersecurity experts in this task which in the current state of the art is entirely manual.
+This paper proposes employing artificial intelligence models as substitutes for humans or as aides to human experts in estimating vulnerabilities.
+We compare the precision, recall, and F1 score amongst the Universal Sentence Encoder, Generative Pre-trained Transformer, and Support Vector Machine, trained on 118,000 vulnerabilities and tested on 51,000 vulnerabilities, with human experts on mean estimation error and variance for each type of vulnerability from the state of the art work in estimating vulnerability severity scores.
+The Universal Sentence Encoder demonstrates superior performance with results (72/77 percent accuracy on sevrity-level prediction) that significantly outperform human experts in assessment tasks for various types of vulnerabilities.
+Additionally, we examine the efficacy of our models in predicting the components and severity level of vulnerabilities.
+The findings highlight the potential of artificial intelligence agents to assist cybersecurity experts in this task which in the current state of the art is entirely manual.

R3.Q2 Regarding the title, I would suggest authors elaborate on the full form of CVSS in the title.

Response We have revised the title to include the full form of CVSS.

> -Comparing Artificial Intelligence Models with Human Experts for Assessing CVSS from Natural Language Descriptions
> +Automated Common Vulnerability Scoring from Common Vulnerabilities and Exposures Descriptions

R3.Q3 Expand on the dataset characteristics, including how it was split into training and test sets.

> +We set the random state to 42 to split the training and test data using the train_test_split function from the SciKit Learn library.

R3.Q4 Include more comprehensive metrics (e.g., precision, recall, F1-score) for each model comparison, not just the final outcome.

Response Pease refer to table 1,2,3,4 which show the comprehensive metrics. For human expert results, since they estimate the base score which is a regression task, precision, recall, f1 aren't applicable. For the ease of comparability, we show the result in the convention from Holm et al.

R3.Q5 Discuss the broader implications of AI outperforming human experts in CVSS scoring. Address any limitations of the current study and potential biases in the dataset or model.

> +When deployed in a real-world scenario, the USE model would be a strong candidate for the CVSS estimation task due to its strong generalization ability and low resource consumption.
> +This research provides a strong argument for the use of AI models in automating the CVSS estimation task to alleviate the issues caused by human errors.
> +A fully automated CVSS estimation pipeline would cut down the time required to process the vulnerabilities and improve the overall security posture of the organization.

R3.Q6 In deep learning, it is widely observed that domain-specific information is very important in classification problems. The choice of models thus does not seem very appropriate to me. The authors have used USE, SVM, and GPT. Why did they not use other transformer-based language models? Those models have seen good performance (Shahid and Debar 2021; Costa et al. 2022). Also, the authors need to use more recent statistics to provide up-to-date context on the significance of vulnerability assessment delays.

Response Deep learning, the using of neural network instead of "shallow learners" like SVM, differs from large language models (LLMs) like GPT-3 in the sense that LLMs are pretrained on a large corpus of text data and fine-tuned on a specific task. We have tried CysecBERT in CVSS estimation which provided marginal increase in performance (0.5-1 percent) in severity level prediction. Since our focus is to compare human expert with an efficient AI model, we decide to focus on USE, which is also transformer based. The pretraining of LLMs could add context information but we suspect if that is beneficial to CVSS estimation task. Our result shows that the

USE model outperforms the GPT-3 model, which represents the state-of-the-art in LLMs. This shows that the current pretraining does not provide a significant advantage in the CVSS estimation task where deep neural networks without pretraining can perform as well with lower running cost.

R3.Q7 The paper misses where the models perform poorly and why. For this reason, as well, authors should also consider using transformer-based language models like BERT, etc. where they can trace the explainability of the model.

Response We have tried BERT with Shap but the results fail to provide any meaningful insight. Therefore, we leave this to future work due to the complexity and the research scope. This paper per se aims to show the potential in automating CVSS scoring, not to mitigate the flaws in the design of CVSS which even experts in the domain have trouble with.

R3.Q8 Please make consistent use of terms and abbreviations throughout the paper. Enhance the readability of figures and tables, ensuring they are self-explanatory. The values in the bar graphs are very hard to read.

Response We have revised the figures and tables to make them more readable. We have also revised use of terms and abbreviations throughout the paper.