- For binary questions, all accuracy appear to be around 0.5, which is at chance level. This raises question about whether the task is really reasonably solved by any of the experimented methods. As mentioned by reviewers, it may be better to conduct an error study to show when the model actually succeeds or fails.
    - If we consider the text-based questions in English language from the original Clotho and mClotho dataset [1], [2] (which contains the same synthesized version of the text in TM-AQA dataset), the accuracy is shown to hover around 62%. Since speech has more variability compared to static text, the best performance of our proposed system dropped to reasonably 55%. Additionally, upon performing qualitative analysis on the original dataset in [1], we find most of the questions were posed multiple times to distinct human subjects and they provided divided answers. For example, in this specific question "*Does the buzzing intensify?*", only one of the subjects has affirmed while the other two negated. When the authors collectively analyzed the audio, we could find that the buzzing sound does intensify and the model has predicted as "Yes" for this binary classification task and is indeed correct but the accuracy will be dropped to one-third or 33.34% for this specific test case. Multiple similar cases can be observed in the original dataset. This collectively brings the overall performance to just above 55 percent accuracy for binary classification tasks. However, the proposed scanning technique has comprehended the environmental sounds better than the existing techniques as presented in our paper. As we have focussed on extending this dataset to cater to speech-based questions in multiple languages, we took the original dataset as it is. But we sincerely would like to thank you for making this extremely important question regarding the qualitative analysis and we realize the gravity. We will include a detailed qualitative analysis as a subsection.

- Reviewers also noticed that the small version of the proposed model outperforms the large one, which is unexpected. I understand this might be due to the limited dataset size, but this also reveals the concern about whether model performances on this dataset really reflect their practical usefulness, when data and models are scaled up.
    - Thank you very much for this important suggestion. We have augmented our dataset by 20 times using various data augmentation techniques (speed perturbation, adding Gaussian noise, audio-shifting, stretching, changing pitch, changing speed, etc) along with feature augmentation techniques (Time and Frequency Masking, Spectrogram Augmentation) and observed the improvement in performance when the data is augmented. Additionally, we also observe that the medium A-MAMBA variant excelled among others for scaled data. Hence this suggests that selection of larger variants of A-MAMBA would be more appropriate when the data size grew. This implies the scalability of our proposed model and practical usefulness when the scalability of the dataset is up. The summary of the results as well as python code can be accessed by the github link [*https://aquorio15.github.io/website/*] and from the table below. Please note

that we have computed for English and Bengali and are undergoing training for Hindi.

English						
Model	**Dataset**	**Variant**	**Binary**	**Multiclass-Top-1**	**Multiclass Top-5**	**Multiclass Top-10**
A-MAMBA + SSM	Augmented dataset	Small	53.87	42.54	69.2	76.56
		Medium	**55.79**	**44.86**	**74.77**	**78.55**
		Large	47.2	36.5	66.8	**73.3**
A-MAMBA + SSM	Original dataset	Small	**52**	**40.59**	**66.87**	68.97
		Medium	51.62	37.26	65.23	66.77
		Large	44.62	33.48	60.85	61.89
A-MAMBA + CSM	Augmented dataset	Small	54.11	44.34	70.23	76.67
		Medium	**58.23**	**45.17**	**77.04**	**79.43**
		Large	50.44	37.12	71.43	76.89
A-MAMBA + CSM	Original dataset	Small	**55.56**	**43.26**	**69.11**	**74.77**
		Medium	51.98	41.42	68.44	71.21
		Large	45.86	35.94	63.8	66.48

Bengali						
Model	**Dataset**	**Variant**	**Binary**	**Multiclass-Top-1**	**Multiclass Top-5**	**Multiclass Top-10**
A-MAMBA + SSM	Augmented dataset	Small	52.39	40.66	68.88	75.3
		Medium	**53.37**	**40.79**	**70.47**	**76.11**
		Large	45.65	37.12	66.9	71.08
A-MAMBA + SSM	Original dataset	Small	**51.21**	**38.42**	**68.75**	**74.32**
		Medium	50.34	36.43	65.87	68.76
		Large	44.89	33.37	62.75	66.76
A-MAMBA + CSM	Augmented dataset	Small	54.04	40.23	70.43	75.96
		Medium	**55.6**	**43.41**	**74.68**	77.22
		Large	46.39	36.32	70.9	**77.88**
A-MAMBA + CSM	Original dataset	Small	**52.78**	**39.44**	**70.11**	**75.76**
		Medium	50.79	38.2	65.78	70.87
		Large	45.78	34.61	63.11	66.88

- The baseline used in the experiments are not competitive enough. Multiple reviewers suggested including a comparison with ASR + text QA model. Although such a model has cons of possible error propagation, it also has clear pros of having stronger models (or even commercial services). Meanwhile, to compare with transformers, a valuable experiment would be to replace MAMBA layers with transformer layers in the proposed model, to verify the performance and computational efficiency benefits of MAMBA as claimed in the paper.
    - For the first part of the question, we are currently training the system and we will hopefully complete but unfortunately we will not be able to throw light at the moment but we will put it as an additional ablation study. In the meantime, the results from [2] can be compared which may serve as the reference but the text translations in multiple languages were not subjected to human evaluation and hence we did not use their text to synthesize.
    - For the second part of the question, we have replaced the MAMBA layers with equivalent transformer layers and evaluated the performance and computational efficiency as follows in the link provided [*https://aquorio15.github.io/website/*] as well as in table below. This table would verify the performance and efficiency of our model for English language for the time-being and we will update in the github link for the remaining two languages too. We will also include this ablation as suggested in the final paper.

English							
# of layers	**Model**	**Params**	**Variant**	**Binary**	**Multiclass Top-1**	**Multiclass Top-5**	**Multiclass Top-10**
1	Transformer	12.7M	Small	46.43	38.06	55.55	61.24
2		19.4M	Medium	51.82	38.66	62.15	66.7
4		48.2M	Large	47.86	38.79	65.77	68.87
1	A-AMAMBA + CSM	11M	Small	55.56	43.26	69.11	74.77
2		18.5M	Medium	51.98	41.42	68.44	71.21
4		45.26M	Large	45.86	35.94	63.8	66.48

- Many reviewers posted concerns about the quality of synthesized speech and whether it reflects model performance in real-world. It is recommended to add more discussions, some data samples, and ideally experiment results on real human speech (acceptable on a smaller scale).
    - Thanks for this suggestion. Currently we have recorded very few samples of test data due to time constraints and have evaluated on those samples and corresponding results in synthetic and original can be found in the link [https://aquorio15.github.io/website/] . However, we are continuing the recording process rapidly using multiple speakers per language and will upload the final results in the link provided and in the final version of the paper.

- In the current version, the introduction of CSM is somewhat unsatisfactory. Some helpful explanations are provided in the rebuttal and could be added into the paper. Basically, beside an example, it would be better to include a rigorous definition, visual illustrations, connections to related work, and more details on the corresponding 2D audio input (specifically, what is the 2D structure of the audio input in A-MAMBA and AST-MAMBA).
    - Thanks for the suggestion regarding illustration of CSS. Please find the illustration in this link [https://aquorio15.github.io/website/] . As suggested, we will add these illustrations in the paper along with an explanation. We hope that the illustrations provided will highlight the differences between other scanning techniques such as Zigzag etc.