Author Responses

Dear Reviewers, we appreciate your positive feedback and insightful comments. In the following, we address the questions of the reviewers point by point.

Reviewer 1

Comment 1: Key Strength: A novel method has been proposed for introducing human-like speech and audio perception task. Main weakness: NA.

Response: We thank the reviewer for the very positive feedback and insightful comprehension of our paper.

Reviewer 2

We thank Reviewer 2 for carefully checking our paper and for the very detailed, insightful, and constructive comments.

Comment 1: More than 50% samples in Open-ASQA are from OpenAQA, i.e. audio, not speech, while speech can be rather diverse in both semantic and paralinguistic sense. ... What if you downsample the data for audio during training to place more focus on speech?

Response: First, the LLaMA LLM already encodes very rich semantic information. So we do not need a huge speech dataset for semantic learning purposes, instead, the main goal is to teach the following the instructions, and the 2.7M data we have should be sufficient. Second, we do need a large speech dataset for paralinguistic learning and we tried our best to include speech datasets with paralinguistic annotations. However, since annotating paralinguistics is expensive, the speech dataset size is still smaller than audio datasets. Performance-wise, in Table 4, Ablation 1, we show two extreme situations that only use audio and speech training datasets, respectively. We show that even using 100% *speech* training data, on *speech* tasks, the performance of the model is still similar to the model trained with the mixed dataset, i.e., downsampling audio data wouldn't lead to a large improvement on speech tasks.

Comment 2: The models are only compared with methods like AudioClip on close-ended tasks, and other LTU variants on open-ended tasks. It could be important to evaluate and compare with other recent LLM-enhanced speech models following the same evaluation protocol...

Response: The main factor that hinders us from making the comparison with these concurrent efforts is 1) these publications are too close to this submission, e.g., Speech-LLaMA is on arXiv on July 8th, while we submit this paper on July 10th; 2) the code or paper is not yet available, e.g., SpeechGPT code is not yet released, and WER of the system has not yet been reported in the paper; ImageBind-LLM releases the code but the paper hasn't been released; 3) tasks and training data have big differences, e.g., SpeechGPT focuses on ASR and cannot recognize paralinguistics; Speech-LLaMA focus on translation. It would not be fair to these papers for audio task benchmarking; 4) Benchmarking multi-modal LLMs is non-trivial and needs a detailed discussion, which would be better to be a separate paper to ensure the soundness and correctness of the comparison. The difference between end-to-end models and text-aligned models is quite obvious in recognizing the details of the sound, e.g., for the question: “Is the keyboard typing loud? Why?” (ESC-50 sample)

LTU-AS: The typing is not particularly loud, but it could be considered moderately loud due to its high-pitched sound.

ImageBind-LLaMA: Yes, the keyboard typing is loud because the user is typing with the ”Loud” key pressed, which likely increases the volume of the keyboard’s typing sounds.

Comment 3: ... it will be interesting to know the roles and importance of different layers in the speech encoder across different tasks, e.g. by ablation studies or showing the attention.

Response: We use the TLTR model proposed in Whisper-AT to capture the information in all layers. The Whisper-AT paper Figure 3 and Table 1 show different sound events need information from different layers, and using the last layer would cause a big performance drop. In addition to audio events, we also need to encode speech paralinguistics in LTU-AS, so it would be necessary to capture information from all Whisper layers. We will clarify this in the next version.

Comment 4: ... I’m curious if the model could also demonstrate some capability understanding non-English speech.

Response: Yes, LTU-AS does understand non-English speech well. We can of course force Whisper to translate non-English speech to English, and then input to LLaMA. However, we found though LLaMA mainly *outputs* English, it does *understand* non-English speech well. We will add a short discussion.

Comment 5: Sec 3.2: How diverse are the generated questions? How different are they compared to each other and the examples in the input prompt?...

Response: The GPT-generated training data are indeed very diverse. Specifically, among the 6.69M open-ended AQAs, 4.36M (65.1%) questions and 6.06M (90.6%) answers are unique, respectively, i.e., most questions are only asked once.

Comment 6: Sec 5.1: ”Even on audio classification and age classification tasks where {S} is not useful” I guess the author means gender instead of age according to Table 4.

Response: Yes, we have fixed this typo.

Comment 7: The authors investigated zero-shot learning capabilities, while a key capability of LLM is few-shot in-context learning. Have the authors attempted to investigate that?

Response: This is a great point. *Few-shot in-context learning* is a key capability of LLM, and `LTU-AS` naturally support this. However, few-shot learning experiments require careful setting (e.g., how many samples to use, how to select samples) and detailed discussion. Due to space limitations, we would leave this as future work and focus on the main topic (joint audio and speech understanding) in this paper.

Comment 8: The manuscripts mentioned that the code and models will be released but didn’t mention the dataset. Will Open-ASQA be released as well?

Response: Open-ASQA dataset will also be released. Specifically, we will release the {question/answer/audio id} tuples. The audio files can be downloaded from public datasets.

Comment 9: There are formatting issues in the references, e.g. [2] should be ”LLaMA” instead of ”Llama”.

Response: We thank the reviewer for pointing this out. We have fixed this and also checked/corrected all other references.

Reviewer 3

Comment 1: A well-written paper. What is required is sufficiently described within 8 pages.

Response: We thank the reviewer for the positive feedback.

Comment 2: The effectiveness of the multi-stage training curriculum is not presented in the experiment section.

Response: The impact of the training curriculum is ablated in the original LTU paper (Table 4), where it was shown that the model performance dramatically drops without the curriculum. LTU-AS uses a similar training framework. Due to space limitations, we skip this in this paper. We will add a short discussion in the next version of the paper.