Evaluating Foundation Models On Timbre-Related Cognitive Tasks

Yorgos Velissaridis **Rajpreet Athwal** **Muneeb Musharaf** **György Fazekas** **Charalampos Saitis**

School of Electronic Engineering and Computer Science, Queen Mary University of London

georgiosvelissaridis@gmail.com

Join my Teams room to discuss live!

Highlights

We evaluated how well AI models capture human responses in *timbre–related* cognitive experiments

Contrastive learning audio–language models (CLAP, MuQ-MuLan) *match* prior state-of-the-art and enable zero-shot inference

Centaur LLM outperforms audio–language models; hybrid CLAP → Centaur pipeline achieves the *closest match* to human responses

Background

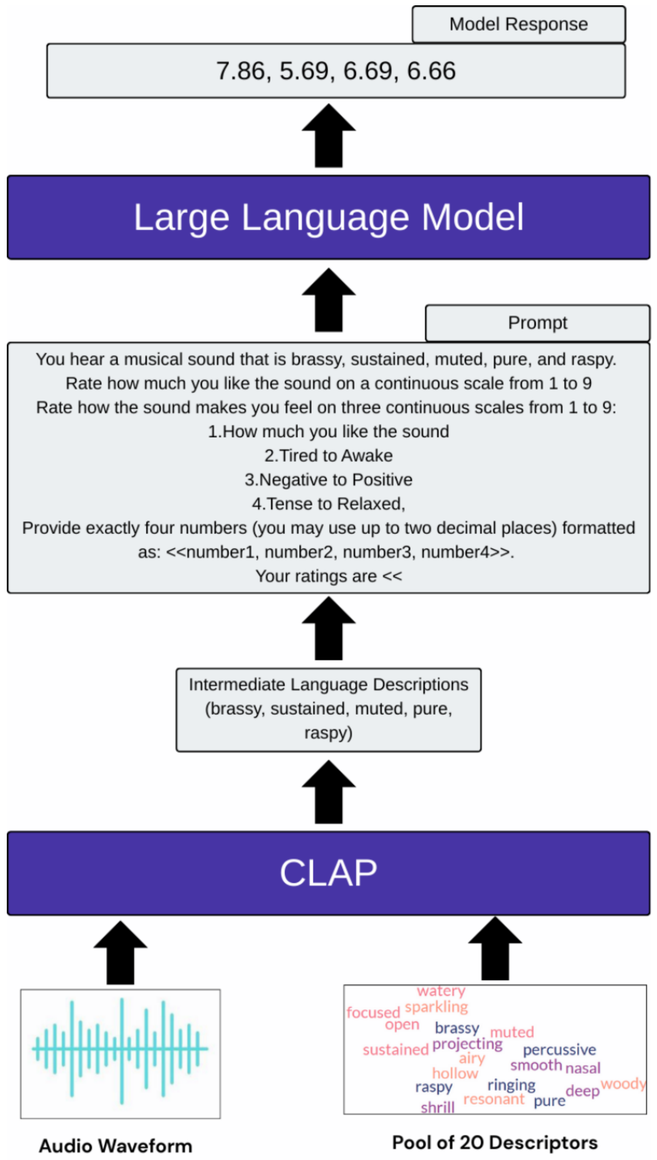
- Music AI has been tested on generation and high level understanding, but not *music cognition*
- Timbre-related behavioural experiments provide a valuable and novel benchmark for foundation models

Research Questions

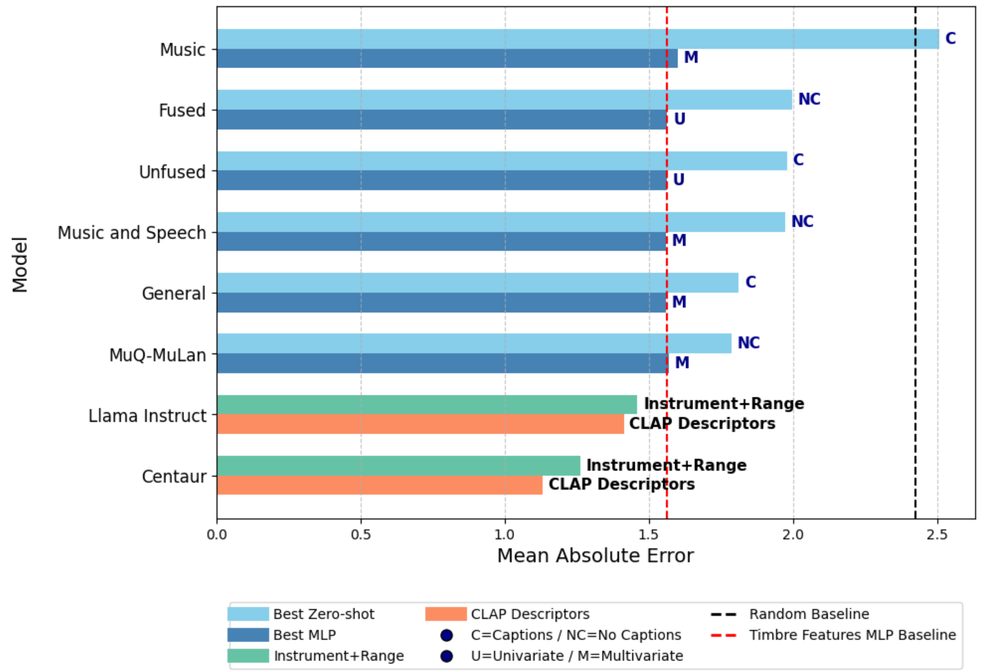
1. How well do foundation models capture responses in timbre emotion and instrument recognition experiments?
2. Is the combination of audio–language models with LLMs better than each on its own?

Data for Timbre Emotion Recognition

Affective descriptors for timbre (Korsmit et al. 2023): 59 tones × 26 instruments, 263 listeners, four sub-experiments (Induced/Perceived × Dimensional/Discrete)



Results: Model Comparison



Conclusions

- Broad pre-training data helps in music cognition tasks
- Centaur outperforms Llama: fine-tuning on human judgements improves performance for music cognition
- Hybrid pipeline combining CLAP model with LLMs show promise as alternatives to end-to-end models

References

Korsmit, I. R., Montrey, M., Wong-Min, A. Y. T., & McAdams, S. (2023). A comparison of dimensional and discrete models for the representation of perceived and induced affect in response to short musical sounds. Frontiers in Psychology, 14, 1287334.
Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., & Dubnov, S. (2023, June). Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.
Zhu, H., Zhou, Y., Chen, H., Yu, J., Ma, Z., Gu, R., ... & Chen, X. (2025). Muq: Self-supervised music representation learning with mel residual vector quantization. arXiv preprint arXiv:2501.01108.
Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., ... & Schulz, E. (2025). A foundation model to predict and capture human cognition. Nature, 1-8.

This work was supported by the UKRI and EPSRC under grant EP/S022694/1

This work was partly carried out as during the Research Ready programme, supported by QMUL and:

György Fazekas, Charalampos Saitis and Yorgos Velissaridis are members of QMUL's