LLaQo: Towards a Query-Based Coach in Expressive Music Performance Assessment

Huan Zhang [*]   Vincent K.M. Cheung[†]   Hayato Nishioka[†]   Simon Dixon[*]   Shinichi Furuya[†]

[*] Centre for Digital Music, Queen Mary University of London, UK

[†] Sony Computer Science Laboratories, Tokyo, Japan

huan.zhang@qmul.ac.uk

Abstract—**Research in music understanding has extensively explored composition-level attributes such as key, genre, and instrumentation through advanced representations, leading to cross-modal applications using large language models. However, aspects of musical performance such as stylistic expression and technique remain underexplored, along with the potential of using large language models to enhance educational outcomes with customized feedback. To bridge this gap, we introduce LLaQo, a Large Language Query-based music coach that leverages audio language modeling to provide detailed and formative assessments of music performances. We also introduce instruction-tuned query-response datasets that cover a variety of performance dimensions from pitch accuracy to articulation, as well as contextual performance understanding (such as difficulty and performance techniques). Utilizing AudioMAE encoder and Vicuna-7b LLM backend, our model achieved state-of-the-art (SOTA) results in predicting teachers' performance ratings, as well as in identifying piece difficulty and playing techniques. Textual responses from LLaQo was moreover rated significantly higher compared to other baseline models in a user study using audio-text matching. Our proposed model can thus provide informative answers to open-ended questions related to musical performance from audio data.**

Index Terms—**large language models, music education, performance assessment, semantical feedback, piano**

I. INTRODUCTION

In the domain of Music Information Retrieval, the majority of music understanding research has focused on composition-level attributes such as key, tempo, genre, and instrumentation. [1]–[3] These attributes are not only tagged individually via end-to-end approaches but have also been the focus of foundation models and unified music representations. Recent advances further extended to cross-modal understanding of music, which enabled tasks such as music captioning [4], [5] and even music reasoning [6] with the help of large language models (LLMs). Despite these technological advances, the nuanced aspects of music performance — such as interpreting complex techniques, recognizing stylistic differences, and evaluating performance quality — remain underexplored [7].

On the other hand, traditional systems for music performance assessment [8] have primarily relied on quantitative metrics that provide summative feedback, often in the form of scores or binary judgments [9]–[11]. Such approaches, while beneficial for certain applications, fall short in educational contexts where formative, detailed and instructional feedback is essential for student development [12], [13]. Meanwhile, performance assessment encompasses multiple dimensions including rhythm accuracy, tone production, and expressive dynamics, which makes semantic descriptors particularly appropriate. Thus, we introduce **LLaQo** (Large Language Query-based Coach), a pioneering approach in the domain of expressive performance assessment that leverages the capabilities of LLMs to analyze and assess music performances at a granular level, providing interpretative guidance for MIR-supported music pedagogy. Constructed with AudioMAE for audio embeddings, Vicuna-7b language backend along

with a Q-former aligner, our model is trained on instruction-tuned data specific to performance assessment. Our contribution[1] can be summarized as follows:

1) We present the first LLM-supported music coach, addressing performance understanding tasks including difficulty analysis, technique analysis, composer recognition, and most importantly, performance assessment.
2) We compiled and instruction-tuned a diverse collection of publicly available performance-understanding datasets, and standardized them into query-response pairs. Additionally, we contribute a newly recorded dataset, *NeuroPiano*, with detailed annotations on 12 performance dimensions.
3) Our model demonstrated superior performance in objective and subjective evaluation tasks over existing audio-language model baselines in performance assessment, as well as difficulty and technique prediction.

II. RELATED WORK

The integration of large language models with audio signal processing has been overviewed in Latif et al. [14] and Wu et al. [15]. Pioneering audio language models (ALM) such as Pengi [16], LTU [17] and SALMONN [18] bridge audio and speech perception with natural language reasoning and specializes in tasks from speech recognition to audio reasoning. Such ALMs are usually composed of a frozen audio encoder (CLAP [19], Whisper [20]) and a frozen causal language model (LLaMA [21], GPT2) as backend, and a trained connection or mapping module that project the input audio embeddings into the text token space. Given an input audio-text pair, the models can generate free-form text to address open-ended tasks and close-ended tasks. In the specific branch of music language understanding, however, research has mainly focused on music context question answering [6] and captioning [4], where models are limited to answering questions on compositional aspects such as genre, mood or instrumentation. On the other hand, pure LLMs showed poor musical understanding and are only restricted to answering questions related to the contextual background behind a musical piece [22], [23]. Models with adequate musical performance understanding are therefore still lacking [7].

Another challenge in the field of music understanding relates to the availability and quality of data. Efforts have been made to generate natural language descriptions from large-scale tag datasets through language models, as seen in music captioning initiatives [5]. Additionally, instruction-tuning has been applied to unify disparate music data annotations for fine-tuning LLMs [6]. In the realm of music education, however, acquiring precise descriptors for performance nuances proves difficult [13], whether through automatic means or manual annotation. This task is further complicated by the inherent

[1]Code is available at: https://github.com/anusfoil/LLaQo, Data is demonstrated in project page: https://bit.ly/3XjUOuX
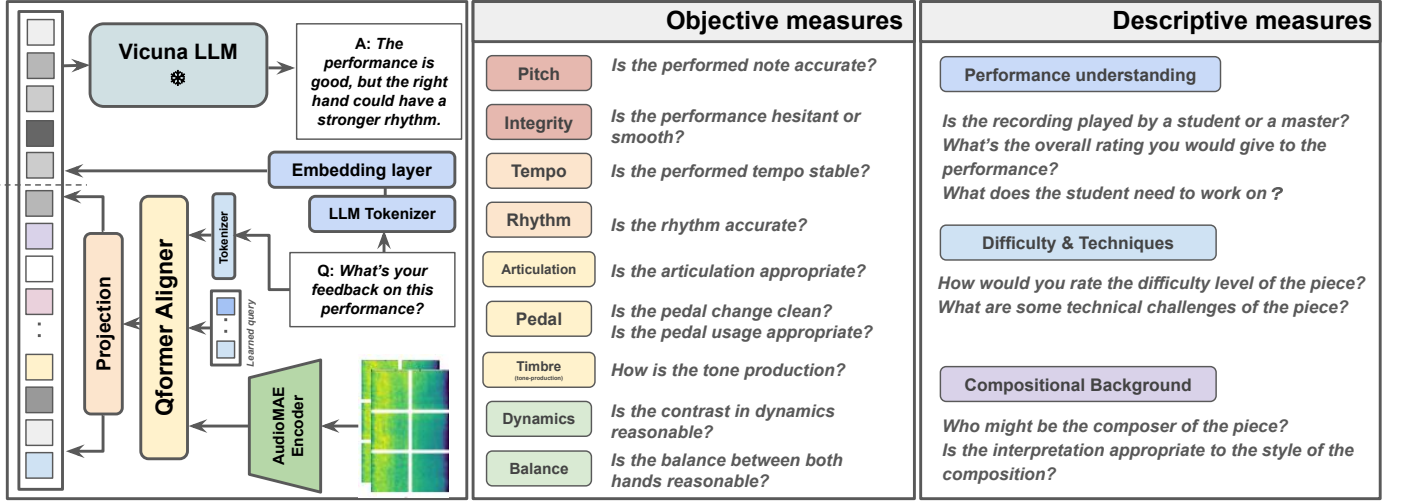
Fig. 1. Overall framework of our model, as well as the piano assessment rubric used to construct query-answer pairs for instruction-tuning in each dataset.

perceptural subjectivity, which can lead to bias and inconsistency in human feedback [8], [24]. Despite these challenges, the field of automatic music assessment has seen considerable benefits from leveraging formative text feedback [12], which provides a richer alternative to traditional methods that rely on simplistic summative scoring systems [9]–[11].

III. INSTRUCTION TUNING DATASET ON MUSIC EDUCATION

In Table I, we describe the instruction-tuned performance datasets used to train our framework. Several types of data were utilized:

The first consisted of teachers' free-response feedback [2] [24], [25] that provides a holistic comment on a performance, covering various performance dimensions or practice suggestions. This data was transformed into query-response pairs using GPT4 according to the instruction of an assessment rubric in Figure 1. For each piece of feedback (around 30 to 50 words), we generated three to five QA pairs.

The second comprised data with performance-related annotations [26], [27], such as a specific articulation or tempo. In this case, queries were formed based on their performance attributes, with examples such as *Is the playing using legato or staccato articulation?*

The third consisted of single-rating datasets [28], [29] that come with a summative rating. The query was designed to ask directly about the rating, such as *How would you rate the overall performance, in a scale of 10?*

The last data type comprised performance understanding data [7], [30] that do not address the quality of a performance, but on contextual information that is ideally obtained by a teacher or student, such as *What's the most difficult technique in this passage?* or *What is the composer and stylistic period of this piece?* In total, we aggregated around 130 hours of recording and 34k QA pairs for training.

For this project, we also recorded a student pianist dataset with detailed teacher annotations (the *NeuroPiano* dataset). It consists of short technique snippets including scales, stepwise chords, and third dyads performed by students[3]. Based on the rubric in Figure 1, we formulated 12 questions, ranging from tone production to hand balance, in which teachers provided textual feedback and a rating score on a scale from 1 to 6. Note that the original annotations are in Japanese, in which we used DeepL to translate to English, and

performed a round of manual correction. The dataset is further spitted by half for training, with the remaining half used for objective and subjective evaluation described in Section V.

IV. MODEL ARCHITECTURE AND TRAINING

The overall structure of LLaQo is based on the APT-LLM [34] framework, which originates from BLIP-2 framework, as shown in Fig 1. It is comprised of an audio encoder, an audio-language aligner, and a large language model. We use Audio-MAE [35], a 12-layer transformer encoder that learns to reconstruct randomly-masked spectrogram patches during training, as the audio encoder. The output feature map from the penultimate block of Audio-MAE encodes fine-grained patterns that's essential for performance understanding.

The audio-language aligner connects the audio encoder to the frozen language model. It takes in a text prompt together with audio feature maps extracted by the audio encoder, and produces a fixed number of acoustic embeddings. Following a Query Transformer (Q-former) architecture, four transformer blocks constitute our audio aligner where 32 trainable query embeddings attend to the input text tokens and extract relevant information from the audio feature maps.

The language model predicts the output text by considering the previous generated texts and the input audio-text tokens. When interleaving audio and text tokens, each audio clip is appended with a learnable token ⟨AUDIO⟩ to indicate the beginning of audio tokens. As shown in Figure 1, the query text pass through both the Q-former tokenizer and LLM's tokenizer and its word embedding layer. We parameterize the language model by Vicuna-7b [36].

A. Learning Objective

Consider an audio-text pair (a, t) and their response g. The audio input a is transformed into a sequence of embeddings $\mathbf{X}_{\text{audio}}$ using the encoder $\mathcal{A}_\phi$ and aligner $\mathcal{M}_\theta$. These embeddings are then concatenated with text embeddings $\mathbf{X}_{\text{text}}$, derived from input text t using the embedding layer $\mathbf{W}_\psi$ to form the combined sequence $\mathbf{X}_{\text{audio;text}}$. More precisely, the embeddings are concatenated as follows:

$$\mathbf{X}_{\text{audio;text}} = C(\mathcal{M}_\theta(\mathcal{A}_\phi(a)), \mathbf{W}_\psi(t)), \quad (1)$$

where C denotes the concatenation function.

The learning objective $\mathcal{L}$ of our model is to optimize the parameters ϕ, θ, and ψ to maximize its ability to predict the next token based on the concatenated embeddings by minimizing cross-entropy loss:

$$\mathcal{L} = \sum_{i=L+1}^{L+|g|} \log p_{\phi,\theta,\psi}(\mathbf{X}_i | \mathbf{X}_{\text{audio;text}}, \mathbf{X}_{\text{pred}, <i}), \quad (2)$$

---

[2]The CROCUS dataset is translated from Japanese to English using DeepL.
[3]From the NeuroPiano Music academy: www.neuropiano.org

Dataset	Recordings	QA pairs	Level	Feedback or Annotation	Repertoire
Expert-Novice ([24])	83	2.4k	beginner	Full verbal feedback and ratings	Pop song arrangements
CROCUS-piano ([25])	22	616	intermediate, advanced	Verbal feedback	Western classical concert repertoire
Con Espressione ([31])	50	4.3k	master, MIDI	Description of expressive character	Western classical concert repertoire
Expressive Musical Gestures ([26])	106	845	intermediate	Labels of performance instruction (e.g. slow, staccato)	'Träumerei' by R. Schumann
Burgmuller ([32])	25	114	intermediate	Annotation of error regions	Burgmuller Etudes
Music Shape Dataset ([27])	2.3K	11k	intermediate	Annotations of dynamics and articulation changes	Schmitt exercise
PISA ([28])	61	256	beginner-intermediate	Single rating score	Folk songs arrangements, easy classical pieces
YCU-PPE ([29])	2.6K	7.7k	beginner	Rating score from 3 judges	Chinese folk song arrangements
PLD-expertise ([7], [33])	1.4k	2k	beginner and advanced	Performing skill level based on the uploader's profile	Western classical concert repertoire
PLD-techniques ([7])	223	223	mixed	Piano Technique labeling (e.g. arpeggio, octave)	Music segments of common piano techniques such as scales, ornaments, repeated notes, etc.
CIPI ([30])	736	2.2k	intermediate to advanced	Henle Difficulty labeling on a scale of 9	Western classical concert repertoire
NeuroPiano (ours)	104	3.3k	advanced	text response on 12 specific aspects including articulation and tempo, as well as rating score on a scale of 6	Short technique snippets including scales, arpeggios, dyads, blocked chords, octaves

TABLE I
PUBLIC DATASETS RELATED TO PERFORMANCE ASSESSMENT AND PERFORMANCE UNDERSTANDING, INSTRUCTION-TUNED FOR TRAINING OUR MODEL.

where L is the length of $\mathbf{X}_{\text{audio;text}}$. This formulation allows us to directly leverage the audio-text context to predict subsequent tokens.

B. Experiment setup

Pretraining: We first pretrain the audio aligner module to bridge the audio modality and the text modality. The Qformer audio aligner is trained with audio-text pairs from AudioSet and WavCaps, with the triplet objective of audio-text matching, audio-grounded text generation, and audio-text contrastive as described in [34]. This step is performed on four NVIDIA A100 (40G) and trained for 120 hours.
General audio multi-task finetuning: The first-stage finetuning trains the whole pipeline as described in the previous section. It is performed on general audio tasks such as *audio tagging, audio captioning, audio question answering,* facilitated by datasets including AudioSet, WavCaps, and Clotho. Furthermore, it learns from multiple audio clips through few-shot audio classification and natural language audio reasoning, where it predicts sound event labels by juxtaposing more than one audio clips with input text. This step is performed on four NVIDIA A100 (40G) and trained for 120 hours.
LLaQo data finetuning: The final round of finetuning is trained with intruction-tuned datasets in Table I, with (a, t, g) triplets as described in Equation 2. This last step is performed on one NVIDIA H100 (80G) GPU and trained for 120 hours.

Adam optimiser was used for all model training described above. We applied linear warmup strategy in the first 2K steps and used a cosine annealing learning rate of 5×10^5. For the AudioMAE input, we resampled all data into 32kHz and computed Kaldi filterbank with 128 mels and a frameshift of $10ms$. Due to resource limitation, our input is limited to a filterbank of length 4096 (40 seconds).

V. EVALUATION AND DISCUSSIONS

A. Compared baselines

For all tasks, we compared our model against two open-source audio-language models capable of answering audio QA tasks, namely **Listen,**

Think and Understand (LTU-AS) [17], which is an improvement of the LTU [37] that incorporates Whisper [20] and low-rank adaptors to the language model, as well as **Music Understanding LLaMa (MuLLaMA)** [4], which is trained on a generated MusicQA dataset and utilizes MERT encoder for audio representation to answer music-related questions and generate captions for tracks.

B. Objective evaluation

Given an audio input, we first evaluated the performance of each model in terms of its ability to answer open and closed questions regarding *Performance assessment*, *Difficulty*, and *Technique*. We performed three iterations of objective evaluation to ensure that the reported results were stable.

To evaluate Performance assessment, each model was required to provide open responses to 12 questions related to the quality of a performance (see Figure 1), as well as an integer score from 1 to 6 (e.g., with the prompt *'How would you rate the consistency of the tempo on a scale of 1 to 6?'*) on the evaluation set of the NeuroPiano dataset. Model performance on open questions was assessed in terms of semantic validity and given in terms of common NLP metrics, namely *BLEU* (B-U) (an average of BLEU1, BLEU2, BLEU3 and BLEU4), *SPICE* [38] and *BERT-Score* (BERT-S). We also measured *sentiment similarity* (SS), which is quantified by the cosine similarity of a 5-class sentiment vector between teachers' feedback and model output as given by a sentiment analysis model[4]. Performance on closed questions was quantified using *mean absolute error* (MAE), which measures the deviation from the predicted rating with the teacher rating, and *prompt following rate* (PFR), which defines the ratio in which the model actually produced a numerical rating. Given that baseline models LTU and Mu-LLaMa are not specifically trained for

---

[4]https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

Model	Performance assessment						Difficulty		Technique
	PFR↑	MAE↓	SS↑	B-U↑	SPICE↑	BERT-S↑	Acc_0 ↑	Acc_1 ↑	Acc_0 ↑
LLaQo	**0.99±0.02**	**0.97±0.10**	**0.52±0.08**	**0.09±0.02**	**0.12±0.04**	**0.88±0.07**	**0.21±0.10**	**0.55±0.15**	**0.37±0.12**
LTU	0.98±0.01	1.39±0.08	0.32±0.05	0.06±0.03	0.07±0.10	0.87±0.12	0.16±0.09	0.34±0.13	0.16±0.07
MU-LLaMa	0.66±0.07	1.81±0.15	0.40±0.10	0.06±0.05	0.12±0.06	0.88±0.09	0.11±0.12	0.26±0.18	0.20±0.10

TABLE II
COMPARISON OF MODELS ON THE METRICS OF THE EVALUATED CONTENT. NUMBERS INDICATE MEAN AND STANDARD DEVIATION. ↑ INDICATES LARGER VALUE FOR BETTER PERFORMANCE, AND VICE VERSA.

numerical predictions, we allowed a maximum of 5 attempts for them to generate a numerical rating.

To evaluate Difficulty and Technique, we followed the same procedure as in [7]. For Difficulty, each model was required to produce an integer score from 1 to 9 given the prompt *'How would you rate the difficulty of the piece on a scale of 9?'* on the CIPI dataset [30]. Model performance was quantified in terms of proportion of exact matches (Acc_0), and proportion of matches that deviate by at most 1 (Acc_1). For Technique, each model was required to select 1 of 7 playing techniques in response to the prompt *'What's the most salient technique used in this recording? Choose from trills, octaves...'* on the PLD-techniques dataset [7], and evaluated in terms of Acc_0.

The results, presented in Table IV-B, reveal that our proposed LLaQo model achieved the best performance amongst all models with regards to all objective metrics. Numerical predictions by LLaQo was within a MAE of less than 1 on a scale of 6, indicating a close alignment between its predictions and the teachers' ground truth. Also, while LLaQo and LTU were able to consistently return numerical ratings, MU-LLaMa struggled, as shown by a low PFR score.

Generated repsonses by LLaQo also showed substantially higher sentiment similarity with the ground truth (0.52) compared to LTU and MU-LLaMA. However, the BLEU score was relatively low across all models compared to typical values of 0.2-0.3 [4] in music captioning tasks. This could be attributed to the expansive semantic space used by the teacher for feedback, where traditional NLP metrics might not have been suited for evaluating feedback responses in this context. Likewise, all models achieved similar BERT-S scores possible because all model responses remained on-topic.

Lastly, while LLaQo surpassed the baseline models for difficulty and technique prediction by a wide margin, it is interesting to note that it did not outperform non-LLM based approaches on the same dataset (e.g. [30] achieved $Acc_0 = 0.39$ in Difficulty, [7] reported $Acc_0 = 0.47$ in Technique).

C. Subjective feedback evaluation

We conducted a human-based evaluation using *audio-text matching* [6] to subjectively evaluate the quality of textual answers provided by each model. Given a query and audio, participants (20 music students and teachers) rated the response from each model on a 4-point scale in terms of their relevance to the question and alignment with the performance content. In particular, three categories of open-ended questions were matched:

1) Feedback: *What's your overall feedback on this performance?*, or *How clean is the attack?*, or *Is the dynamics change natural?*
2) Suggestion: *What does the student need to work on?*
3) Appreciation: *How would you describe the emotional intent of the performance?*

Participants rated 250 randomly sampled questions on Feedback and Suggestion from the *NeuroPiano* dataset, as well as 250 questions on Feedback, Suggestion, and Appreciation from the Expert-Novice, CROCUS and YCU-PPE datasets.
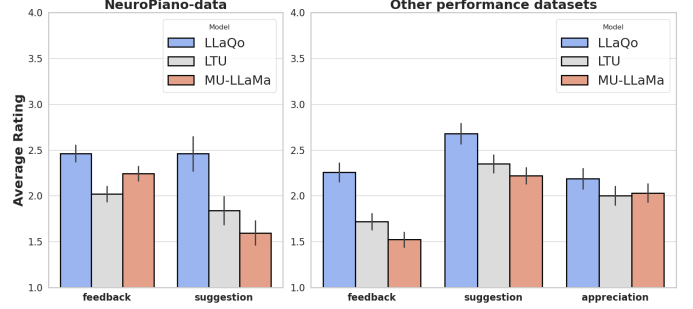

Fig. 2. Audio-text matching subjective evaluation result of three models, with errors bars indicating standard error of the mean.

Results of the audio-text matching task are shown in Figure 2. In terms of Suggestions, participants rated our proposed LLaQo model significantly higher than LTU and MU-LLaMa in both *NeuroPiano* and Other datasets (Mann-Whitney U-test, *NeuroPiano*: $p(corrected) = 0.0274$ and $p = 0.0013$, respectively; Other: $p = 0.0382$ and $p = 0.0028$, respectively). Similarly, LLaQo was rated the highest for Feedback in both datasets, scoring significantly higher than LTU in both ($p = 7.2 \times 10^{-4}$ and $p = 1.9 \times 10{-4}$, respectively) and MU-LLaMa in Others ($p = 0.0028$). Lastly, participants gave highest ratings for LLaQo in questions related to Appreciation, though the differences were not statistically significant.

VI. CONCLUSION AND FUTURE WORK

This paper introduced the LLaQo model, designed for music-education related question answering, which is the first study that use Large Language Model to support music pedagogy. Our model offers performance feedback and suggestions, as well as information on performance techniques and piece difficulty, and has demonstrated superior performance in objective and subjective evaluation tasks over existing models such as LTU and MU-LLaMa. Additionally, we contributed a collection of instruction-tuned music question answering datasets, including the recorded and annotated in-house, and previously unreleased *NeuroPiano* dataset. Though our model is currently trained on piano sets, our framework can be easily extended towards other instruments, when large-scale assessment data become available. Future work should also consider timestamp positioning of feedback, as the ability to correctly locate an error, (e.g. the first measure, the last beat) is crucial for informative feedback.

REFERENCES

[1] F. Korzeniowski and G. Widmer, "End-to-end musical key estimation using a convolutional neural network," in *25th European Signal Processing Conference, EUSIPCO*, 2017.

[2] J. Pons, O. Nieto, M. Prockup, E. Schmidt, A. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 637–644, 2018.

[3] H. Flores Garcia, A. Aguilar, E. Manilow, and B. Pardo, "Leveraging hierarchical structures for few-shot musical instrument recognition," in *Proceedings of the 22th International Society for Music Information Retrieval Conference, ISMIR 2021*, 2021.

[4] S. Liu, A. S. Hussain, C. Sun, and Y. Shan, "Music understanding LLaMA: Advancing text-to-music generation with question answering and captioning," in *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[5] S. Doh, K. Choi, J. Lee, and J. Nam, "LP-MusicCaps: LLM-based pseudo music captioning," in *Proceeding of the 24th International Society on Music Information Retrieval (ISMIR)*, 2023.

[6] J. Gardner, S. Durand, D. Stoller, and R. M. Bittner, "Llark: A multi-modal foundation model for music," in *In Proceedings of International Conference on Machine Learning (ICML)*, Oct. 2024.

[7] H. Zhang, J. Liang, and S. Dixon, "From audio encoders to piano judges: Benchmarking performance understanding for solo piano," in *Proceeding of the 25t International Society on Music Information Retrieval (ISMIR)*, 2024.

[8] A. Lerch, C. Arthur, A. Pati, and S. Gururani, "Music performance analysis: A survey," in *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[9] H. Zhang, Y. Jiang, T. Jiang, and P. Hu, "Learn by referencing: Towards deep metric learning for singing assessment," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[10] J. Huang, Y.-N. Hung, A. Pati, S. K. Gururani, and A. Lerch, "Score-informed networks for music performance assessment," in *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[11] P. Seshadri and A. Lerch, "Improving music performance assessment with contrastive learning," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[12] V. Eremenko, A. Morsi, J. Narang, and X. Serra, "Performance assessment technologies for the support of musical instrument learning," *Proceedings of the 12th International Conference on Computer Supported Education (CSME)*, 2020.

[13] A. Morsi, H. Zhang, A. Maezawa, S. Dixon, and X. Serra, "Simulating piano performance mistakes for music learning," *In Proceedings of the Sound and Music Computing Conference (SMC)*, 2024.

[14] S. Latif, M. Shoukat, F. Shamshad, M. Usama, Y. Ren, H. Cuayáhuitl, W. Wang, X. Zhang, R. Togneri, E. Cambria, and B. W. Schuller, "Sparks of large audio models: A survey and outlook," pp. 1–32, 2023.

[15] H. Wu, X. Chen, Y.-C. Lin, K.-w. Chang, H.-L. Chung, A. H. Liu, and H.-y. Lee, "Towards audio language modeling - an overview," pp. 4–8, 2024.

[16] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, "Pengi: An audio language model for audio tasks," in *In Proceedings of the Conference on Neural Information Processing Systems*, 2023.

[17] Y. Gong, A. H. Liu, H. Luo, L. Karlinsky, and J. Glass, "Joint audio and speech understanding," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU*, 2023.

[18] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Salmonn: Towards generic hearing abilities for large language models," in *Proceedings of the International Conference on Learning Representations (ICLR)*, vol. 82, no. 2, 2024.

[19] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2023-June, 2023.

[20] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of Machine Learning Research*, vol. 202, 2023, pp. 28 492–28 518.

[21] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-a. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," *arXiv:2302.13971*, 2023.

[22] J. Li, L. Yang, M. Tang, C. Chen, Z. Li, P. Wang, and H. Zhao, "The music maestro or the musically challenged , a massive music evaluation benchmark for large language models," in *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[23] Y. Ma, A. Øland, A. Ragni, B. M. Del Sette, C. Saitis, C. Donahue, C. Lin, C. Plachouras, E. Benetos, E. Quinton, E. Shatri, F. Morreale, G. Zhang, G. Fazekas, G. Xia, H. Zhang, I. Manco, J. Huang, J. Guinot, L. Lin, L. Marinelli, M. W. Y. Lam, M. Sharma, Q. Kong, R. B. Dannenberg, R. Yuan, S. Wu, S.-L. Wu, S. Dai, S. Lei, S. Kang, S. Dixon, W. Chen, W. Huang, X. Du, X. Qu, X. Tan, Y. Li, Z. Tian, Z. Wu, Z. Wu, Z. Ma, and Z. Wang, "Foundation models for music: A survey," *arXiv:2408.14340*, 2024.

[24] Y. Jiang, "Expert and novice evaluations of piano performances : Criteria for computer-aided feedback," in *Proceeding of the 24th International Society on Music Information Retrieval (ISMIR)*, 2023.

[25] M. Matsubara, R. Kagawa, T. Hirano, and I. Tsuji, "CROCUS: Dataset of musical performance critiques," in *In Proceedings of the International Symposium on Computer Music Multidisciplinary Research (CMMR)*, 2021.

[26] A. Sarasuá, B. Caramiaux, A. Tanaka, and M. Ortiz, "Datasets for the analysis of expressive musical gestures," in *In Proceedings of the 4th International Conference on Movement Computing (MOCO)*, 2017.

[27] X. Li, S. Weiss, Y. Yan, Y. Li, J. Ren, J. Soraghan, and M. Gong, "Siamese residual neural network for musical shape evaluation in piano performance assessment," in *Proceeding of the 31st European Signal Processing Conference*, 2023, pp. 1–5.

[28] P. Parmar, J. Reddy, and B. Morris, "Piano skills assessment," in *IEEE 23th International Workshop on Multimedia Signal Processing (MMSP)*, 2021.

[29] W. Wang, J. Pan, H. Yi, Z. Song, and M. Li, "Audio-based piano performance evaluation for beginners with convolutional neural network and attention mechanism," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 29, pp. 1119–1133, 2021.

[30] P. Ramoneda, D. Jeong, V. Eremenko, N. C. Tamer, M. Miron, and X. Serra, "Combining piano performance dimensions for score difficulty classification," *Expert Systems with Applications*, 2024.

[31] C. Cancino-Chacón, S. Peter, S. Chowdhury, A. Aljanaki, and G. Widmer, "On the characterization of expressive performance in classical music: First results of the con espressione game," in *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[32] A. Morsi, K. Tatsumi, A. Maezawa, T. Fujishima, and X. Serra, "Sounds Out of Pläce? Score-independent detection of conspicuous mistakes in piano performances," in *Proceeding of the 24th International Society on Music Information Retrieval (ISMIR)*, 2023.

[33] H. Zhang, J. Tang, S. Rafee, S. Dixon, and G. Fazekas, "ATEPP: A dataset of automatically transcribed expressive piano performance," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Bengaluru, India, 2022.

[34] J. Liang, X. Liu, W. Wang, M. D. Plumbley, H. Phan, and E. Benetos, "Acoustic prompt tuning: Empowering large language models with audition capabilities," *arXiv: 2312.00249*, 2023.

[35] P.-y. H. Hu, X. Juncheng, C. Feichtenhofer, and M. Ai, "Masked autoencoders that listen," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[36] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," in *NeurIPS 2023 Datasets and Benchmarks Poster*.

[37] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. Glass, "Listen, think, and understand," in *In proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[38] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *In Proceedings of the European Conference on Computer Vision*, 2016.