Fine-Tuning LLMs for Multi-Track Audio Analysis: A Framework for Music Mixing Assistance

Pratham Vadhulas and Alexander Lerch

INTRODUCTION

Current Audio-Language Models (ALMs) show promise for descriptive audio tasks, but effective music mixing requires a relational understanding of multi-track audio that goes beyond simple analysis. This research investigates a novel framework where an ALM is conditioned on a designated anchor track to learn the relative levels of other stems and generate actionable gain-balancing advice. The study will leverage the MUSDB18 dataset for training and evaluation, combining automated metrics with qualitative assessments from audio engineers to gauge the musicality, effectiveness, and real-world usefulness of the generated advice.

Research Questions

Primary Research Question: To what extent can an Audio Language Model, conditioned on an anchor track, learn about relative levels of multi-track audio to generate effective gain-balancing advice for music mixing?

Secondary Research Questions:

1) **Architecture** What is an effective model architecture for representing multi-track stems and anchor tracks, enabling an Audio Language Model to learn and reason about their relative levels for music mixing?
2) **Learned Conventions** Does the model's generated advice demonstrate an understanding of established mixing conventions and genre-specific expectations (e.g., the typical vocal-to-instrument balance in pop versus jazz)?
3) **Evaluation & Usefulness** How do audio engineers and producers rate the effectiveness, musicality, and actionability of the generated gain-balancing advice, and what qualitative feedback do they provide on its integration into their workflow?
4) **Metric Correlation** What is the correlation between subjective human preference judgments and automated evaluation metrics (e.g., LLM-as-a-Judge, BERTscore) for evaluating mixing advice?

Scope and Limitations

The model's output will be advisory text that may include gain predictions, but evaluation will focus on the overall quality and usefulness of the advice rather than the accuracy of specific numerical predictions. The investigation excludes other mixing parameters such as equalization (EQ), dynamic range compression, and spatial effects. The proposed system is designed for offline analysis and is not intended for real-time, interactive applications. The validity of this work relies on the assumption that the professionally mixed versions of the tracks in the MUSDB18 dataset represent a perceptually valid ground truth for a well-balanced mix and that the dataset is of sufficient quality and diversity for the task.

MOTIVATION

Research Gap

Current automatic mixing systems operate as "black boxes," generating final outputs without providing explainable, actionable guidance that audio engineers need. No existing systems provide mixing advice based on multi-track analysis and relative gain relationships, leaving a gap in the ability to learn and reason about the nuanced, context-dependent gain relationships that are fundamental to professional mixing.

Research Impact

This research aims to bridge this gap by teaching the model to reason about mix balance relative to a stable reference point (anchor track), aligning its understanding with the cognitive workflows of human engineers. This approach enables human-AI collaboration by providing musicians and producers with contextually relevant, actionable advice that can integrate into their existing creative workflow, supporting a more interactive and transparent mixing process.

RELATED WORK

The field of "Co-Creative AI" provides a compelling framework for music production, envisioning AI not as a replacement for human creativity, but as an intelligent partner that augments the artistic workflow [1].

Audio Mixing Paradigms

Traditional automated mixing approaches have evolved through several paradigms: Rule-based methods employed expert knowledge systems [2]–[4], while recent black box systems take raw multitrack audio and produce fully mixed outputs with minimal user intervention [5], [6]. However, these approaches limit the fine-grained control required in professional workflows.

In contrast, the emerging Language Bridge paradigm uses natural language as the primary interface for interaction, allowing users to articulate creative intent in descriptive terms. Recent work has begun exploring this direction [7]–[13].

Audio-Language Models

The development of audio-language models has progressed from task-specific supervised models to more sophisticated approaches. Contrastive dual-encoder models like CLAP learn audio concepts from natural language supervision, enabling zero-shot audio classification and retrieval [14], [15]. Large Audio-Language Models (LALMs) represent a significant advancement, with models demonstrating language modeling approaches to audio generation [16], [17].

Multimodal Large Language Models have emerged with different architectural approaches, including modality interface architectures that augment pre-trained LLMs with specialized audio encoders [18], [19], and any-to-any systems that enable both perception and generation across multiple modalities [20], [21].

LLM Integration Approaches

The integration of LLMs in audio tasks follows several key patterns [22]:

- LLMs as Backbone: Pre-trained LLMs serve as the central architecture with modality-specific encoders/decoders
- LLMs as Conditioner: LLMs encode text prompts into embeddings that condition audio generation
- LLMs as Agent: LLMs act as controllers, coordinating external tools to accomplish audio tasks

Our research builds upon this foundation, specifically targeting the Language Bridge paradigm for audio mixing applications.

PROPOSED METHOD

This research presents a framework for fine-tuning an Audio-Language Model (ALM) to generate gain-balancing advice for music mixing. The approach conditions the model on a designated anchor track to learn relational understanding of multi-track audio levels.

Data Preprocessing and Dataset Creation

We utilize the MUSDB18HQ dataset [23] to synthesize a supervised fine-tuning (SFT) dataset in JSONL format. Our pipeline standardizes chunking, gating, anchor selection, error labeling, and instruction/response generation.

a) Chunking and gating.: Each song is segmented into non-overlapping 10 s chunks. We gate out silent or near-silent segments using empirically tuned thresholds derived from a prior corpus analysis: mixture RMS gate at -25 dBFS, per-stem activity gate at -40 dBFS with a minimum active-frame ratio of 0.30 (50 ms frames). Paths are stored relatively (e.g., `data/...`) to avoid duplicating audio.

b) Anchor selection.: For every chunk we select an anchor stem to establish a reference for relative level reasoning. The policy prefers *bass* when active; otherwise falls back to *drums→vocals→other* using the same activity gates.

c) Track-wide, IQR-scaled error labeling.: Since perceived "quiet/loud" is track-dependent, we adapt error magnitudes per track. For each track we pool RMS dBFS across stems and compute the median and IQR. For a single non-anchor target stem per chunk, we sample an error category with priors (`no_error`, `quiet`, `very_quiet`, `loud`, `very_loud`) and set a target level as median $\pm\ \alpha \cdot \mathrm{IQR}$ (e.g., 0.75 for `quiet/loud`, 1.5 for `very_quiet/very_loud`). The intended gain (in dB) is the difference between the target and the current per-chunk stem level, clamped to $[-12, 12]$ dB. Labels are stored as `error_labels.jsonl`.

d) Instruction and response synthesis.: For this project we focus on **audio-only conditioning** with constant instruction styles and response-only supervision. Each sample includes:

- An **instruction** drawn from a small set of templates that describes the current chunk context (duration, available stems, anchor stem), without revealing the injected error.
- A **response** drawn from category-specific templates that advise a gain change for the target stem (e.g., "Reduce drums by 3 dB"), parameterized by the adaptive intended gain.

Although we only use audio-context instructions here, this design is extensible to additional instruction categories (e.g., advice-seeking, correct/incorrect descriptions) in future work.

e) Training-time flawed mixture.: To create the input audio, we apply the intended gain to the labeled target stem for the chunk and sum stems to produce a *flawed mixture*. The model is trained to generate the corrective response given the flawed audio and the instruction context.

f) Processing pipeline.: The end-to-end pipeline is scripted as: (1) corpus statistics and threshold tuning; (2) chunking with gating and anchor selection; (3) per-track stats aggregation (median/IQR); (4) IQR-scaled error labeling; (5) instruction/response synthesis to JSONL.

Model Architecture and Training Strategy

We will use Qwen-Audio [18] as the base model, while exploring other pre-trained Audio-Language Models for fine-tuning. We will use Low-Rank Adaptation (LoRA) [24] and QLoRA [25] for parameter-efficient fine-tuning.

Data Augmentation and Cross-Sample Training

We will implement cross-sample training strategies to improve generalization. We will explore creating hybrid training examples by combining stems from different songs while maintaining musical coherence.

PROPOSED EVALUATION

The evaluation will assess the model's performance through human evaluation and automated metrics to address the research questions about effectiveness, learned conventions, and metric correlation.

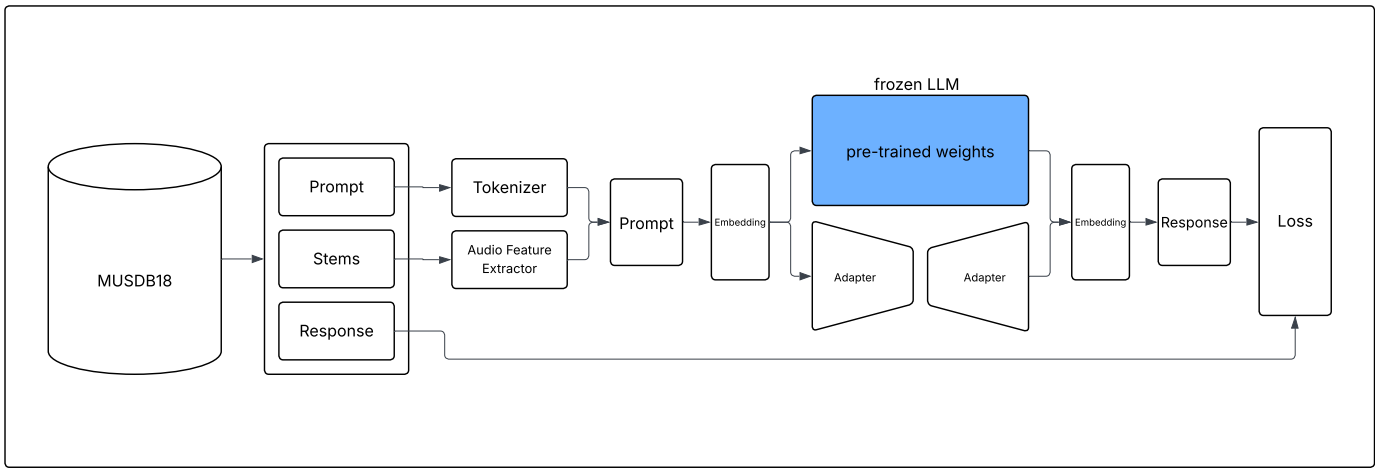

Fig. 1. Proposed architecture of Audio-Language Model fine-tuning for music mixing advice generation.

Human Evaluation

We will conduct a study with semi-professional audio engineers to evaluate the generated mixing advice. They will evaluate mixing scenarios consisting of multi-track stems with designated anchor tracks and gain imbalances, along with the model's generated advice. Participants will rate the advice on a Likert scale across five dimensions:

- Effectiveness: How well the advice addresses the mixing problem
- Musicality: Appropriateness for the genre and musical context
- Conventions: Understanding of established mixing conventions
- Actionability: Clarity and ease of implementation
- Helpfulness: Real-world workflow utility

Automated Evaluation

We will compute automated metrics to complement the human evaluation. We will use LLM-as-a-Judge (GPT-4 evaluation with rubric) and semantic similarity metrics (BERTScore, ROUGE-L, METEOR). For gain advice accuracy, we will measure direction accuracy (whether the advice correctly identifies if a track should be louder or quieter) and intensity accuracy (how well the suggested gain intensity changes match the optimal adjustments). We will analyze the correlation between human ratings and automated metrics to understand how well objective measures align with human perception of mixing advice quality.

DELIVERABLES

Research Publications

- **Poster Presentation**: Presentation of results at 7100 final presentations
- **7100 Paper Submission**: Final research paper submission for course requirements
- **Conference Paper Submission**: Submission to a conference (ISMIR, ICASSP, or AES)

Software and Code

- **Public Codebase**: Open-source implementation with user-friendly documentation and setup instructions
- **Web Interface**: Browser-based interface for testing model deployment and generating mixing advice

Datasets and Models

- **Hugging Face Repository**: Public release of the chat dataset and trained model on Hugging Face for community access and reproducibility

TIMELINE

Current Progress (Completed)

- **Dataset Setup**: MUSDB18 preprocessing and JSONL format conversion completed
- **Codebase and Pipeline Setup**: Initial implementation and data loading pipeline completed
- **Partial Experiments**: Some fine-tuning experiments and architecture testing completed

Remaining Tasks (Leading to November 28th Submission)

- **CITI Certification**: Complete CITI training for IRB approval
- **Experiment Definition and Execution**: Finalize all experimental designs and complete remaining fine-tuning experiments
- **IRB Approval**: Submit and obtain approval for human evaluation studies
- **Human Evaluation**: Conduct evaluation study with audio professionals (anticipated longest phase due to human involvement)
- **Final Deliverables**: Complete paper writing, develop web interface, and deploy to Hugging Face

Future Expansion (Post-Milestone 0)

This milestone serves as a foundation for potential expansion in three directions:

- **Differentiable Tools**: Integration of differentiable audio processing tools for end-to-end training
- **Extended Audio Effects**: Expansion to include additional mixing parameters (EQ, compression, spatial effects) using the same anchor-conditioned architecture
- **Dataset Quality Assessment & Curating**: Study whether current multi-track datasets meet professional mixing standards and curate a new dataset with expert consensus to ensure realistic and informed training data.

Risk Mitigation

- **Human Evaluation Delays**: Begin IRB process early and maintain flexible recruitment strategies
- **Timeline Buffer**: Web interface and Hugging Face deployment can be simplified or delayed. Conference paper submission can be later than november 28th. Core research paper (7100 submission) and poster presentation are fixed requirements.

REFERENCES

[1] R. Louie, A. Coenen, C. Z. Huang, M. Terry, and C. J. Cai, "Novice-ai music co-creation via ai-steering tools for deep generative models," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Honolulu HI USA: ACM, Apr. 2020, p. 1–13. [Online]. Available: https://dl.acm.org/doi/10.1145/3313831.3376739

[2] D. E. Dugan, "Automatic microphone mixer," Dec. 1976, uS Patent 3,992,584. [Online]. Available: https://patents.google.com/patent/US3992584A

[3] E. Pérez-González and J. D. Reiss, "A knowledge-engineered autonomous mixing system," in *Audio Engineering Society Convention 135*, Oct. 2013. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=16953

[4] S. Mansbridge, S. Finn, and J. D. Reiss, "Implementation and evaluation of autonomous multi-track fader control," *Journal of the Audio Engineering Society*, vol. 60, no. 10, pp. 821–839, 2012.

[5] C. J. Steinmetz, J. Pons, S. Pascual, and J. Serrà, "Automatic multitrack mixing with a differentiable mixing console of neural audio effects," no. arXiv:2010.10291, Oct. 2020, arXiv:2010.10291 [eess]. [Online]. Available: http://arxiv.org/abs/2010.10291

[6] M. A. Martínez-Ramírez, W.-H. Liao, G. Fabbro, S. Uhlich, C. Nagashima, and Y. Mitsufuji, "Automatic music mixing with deep learning and out-of-domain data," no. arXiv:2208.11428, Aug. 2022, arXiv:2208.11428 [eess]. [Online]. Available: http://arxiv.org/abs/2208.11428

[7] M. P. Clemens and A. Marasovic, "Mixassist: An audio-language dataset for co-creative AI assistance in music mixing," 2025. [Online]. Available: https://openreview.net/forum?id=5mICyyD4OF

[8] A. Chu, P. O'Reilly, J. Barnett, and B. Pardo, "Text2fx: Harnessing clap embeddings for text-guided audio effects," no. arXiv:2409.18847, Feb. 2025, arXiv:2409.18847 [eess]. [Online]. Available: http://arxiv.org/abs/2409.18847

[9] S. Doh, J. Koo, M. A. Martínez-Ramírez, W.-H. Liao, J. Nam, and Y. Mitsufuji, "Can large language models predict audio effects parameters from natural language?" no. arXiv:2505.20770, Jul. 2025, arXiv:2505.20770 [cs]. [Online]. Available: http://arxiv.org/abs/2505.20770

[10] J. Melechovsky, A. Mehrish, and D. Herremans, "Sonicmaster: Towards controllable all-in-one music restoration and mastering," no. arXiv:2508.03448, Aug. 2025, arXiv:2508.03448 [eess]. [Online]. Available: http://arxiv.org/abs/2508.03448

[11] S.-C. Lai, Y.-H. Hung, Y.-C. Zhu, S.-T. Wang, M.-H. Sheu, and W.-H. Juang, "A low-cost smart digital mixer system based on speech recognition," *Electronics*, vol. 11, no. 4, p. 604, Feb. 2022.

[12] B. Pardo, M. Cartwright, P. Seetharaman, and B. Kim, "Learning to build natural audio production interfaces," *Arts*, vol. 8, no. 3, p. 110, Aug. 2019.

[13] S. Venkatesh, D. Moffat, and E. R. Miranda, "Word embeddings for automatic equalization in audio mixing," *Journal of the Audio Engineering Society*, vol. 70, no. 9, p. 753–763, Nov. 2022.

[14] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023, p. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/10095889

[15] A. S. Koepke, A.-M. Oncescu, J. F. Henriques, Z. Akata, and S. Albanie, "Audio retrieval with natural language queries: A benchmark study," *IEEE Transactions on Multimedia*, vol. 25, p. 2675–2685, 2023.

[16] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, and N. Zeghidour, "Audiolm: A language modeling approach to audio generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, p. 2523–2533, 2023.

[17] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[18] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," no. arXiv:2311.07919, Dec. 2023, arXiv:2311.07919 [eess]. [Online]. Available: http://arxiv.org/abs/2311.07919

[19] 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," no. arXiv:2310.13289, Apr. 2024, arXiv:2310.13289 [cs]. [Online]. Available: http://arxiv.org/abs/2310.13289

[20] A. Vyas, B. Shi, M. Le, A. Tjandra, Y.-C. Wu, B. Guo, J. Zhang, X. Zhang, R. Adkins, W. Ngan, J. Wang, I. Cruz, B. Akula, A. Akinyemi, B. Ellis, R. Moritz, Y. Yungster, A. Rakotoarison, L. Tan, C. Summers, C. Wood, J. Lane, M. Williamson, and W.-N. Hsu, "Audiobox: Unified audio generation with natural language prompts," no. arXiv:2312.15821, Dec. 2023, arXiv:2312.15821 [cs]. [Online]. Available: http://arxiv.org/abs/2312.15821

[21] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, "NExt-GPT: Any-to-any multimodal LLM," 2024. [Online]. Available: https://openreview.net/forum?id=0A5o6dCKeK

[22] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *National Science Review*, vol. 11, no. 12, p. nwae403, Nov. 2024.

[23] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "Musdb18-hq - an uncompressed version of musdb18," Aug. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3338373

[24] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," no. arXiv:2106.09685, Oct. 2021, arXiv:2106.09685 [cs]. [Online]. Available: http://arxiv.org/abs/2106.09685

[25] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *arXiv preprint arXiv:2305.14314*, 2023, arXiv:2305.14314. [Online]. Available: https://arxiv.org/abs/2305.14314