What Matters for Music-Centered Recognition in Audio-Language Models?

Wenye Ma [1] **Ichiro Fujinaga** [1]

Abstract

Audio-language models (AudioLMs) offer a flexible natural-language interface for open-ended audio understanding, yet their efficacy on structured, closed-set audio recognition remains poorly understood. This paper systematically investigates this question across four diverse recognition tasks, with a primary focus on music-centered domains. Internal probing shows that task-relevant information often remains accessible after the audio encoder, while controlled comparisons show that LLM-based label generation changes recognition behavior in a task-dependent way. Specifically, the language interface consistently benefits environmental sound recognition but degrades music genre tagging compared to direct encoder classification. Systematic evaluations of model design choices further establish small-learning-rate encoder tuning as the most robust adaptation strategy, while attention-based multi-encoder fusion gives the strongest in-domain joint model. These results suggest that closed-set AudioLM recognition is limited less by a simple loss of audio information than by how audio representations are adapted, combined, and decoded into fixed labels.

1. Introduction

Audio-language models (AudioLMs) connect audio encoders to large language models (LLMs), giving audio systems a flexible natural-language interface. This interface is appealing for open-ended audio and music queries, but fluent language generation can obscure what the model has actually learned from the audio signal. For music-centered understanding, an AudioLM should not only produce plausible text; its responses should be grounded in reliable audio and music recognition. This raises a basic question: when the target output is a fixed MIR-style label set, does the LLM interface improve recognition, or does it mainly repackage audio evidence through language decoding?

This question matters because many MIR tasks, such as instrument recognition, key detection, and genre tagging, require stable decisions over well-defined label spaces rather than open-ended dialogue. In such settings, generation-based evaluation can conflate audio representation quality with language priors, prompt following, and label parsing. Similar concerns have motivated controlled studies in vision-language modeling, where adding an LLM does not automatically improve closed-set recognition and performance often depends on sensory encoder and connector design (Cooper et al., 2025; Tong et al., 2024). We ask whether the same issue arises for AudioLMs, particularly in music-centered recognition.

We study four structured audio-recognition tasks: instrument recognition, key detection, genre tagging, and environmental sound recognition. The first three are music-centered tasks, while environmental sound recognition serves as a non-music control. All tasks use fixed label vocabularies, allowing us to evaluate both internal probes and generated labels against the same canonical label space and task metrics.

We organize the study around three questions, summarized in Figure 1. First, is task information lost as audio representations pass through existing AudioLM pipelines? To answer this, we probe existing AudioLMs at the encoder, connector, and LLM audio-token stages. Second, when the audio encoder is fixed, does LLM-based label generation outperform direct encoder classification? We test this with controlled AudioLM-style models built on a shared Qwen2.5-7B-Instruct backbone (Qwen et al., 2025). Third, if the LLM interface is not uniformly beneficial, which encoder and connector design choices improve joint recognition? We evaluate encoder tuning, caption-based connector pretraining, and multi-encoder aggregation under joint four-task training.

Our results show that closed-set AudioLM recognition cannot be explained by a simple information-loss story. Existing AudioLMs often retain task-relevant information beyond the audio encoder, although the most informative stage depends on the model and task. However, controlled com-

[1]Schulich School of Music, McGill University, Montreal, QC, Canada. Correspondence to: Wenye Ma <wenye.ma@mail.mcgill.ca>.

Accepted to the ICML 2026 Workshop on Machine Learning for Audio, Seoul, South Korea, 2026.

parisons show that LLM-based label generation is not a uniformly stronger recognition head: it is most helpful for environmental sound recognition, consistently weaker for genre tagging, and mixed for instrument and key recognition. These task-dependent patterns motivate a closer look at encoder and connector design, where small-learning-rate encoder tuning provides the most reliable adaptation strategy and attention-based multi-encoder aggregation gives the strongest in-domain joint model.

2. Tasks, Data, and Evaluation Protocol

We evaluate four structured audio-recognition tasks. For each task, models are trained on one source dataset and evaluated on two test settings: an in-domain test split from the same dataset and a cross-dataset test set with a mapped label vocabulary. Instrument recognition uses OpenMIC-2018 (Humphrey et al., 2018) as the source dataset and MTG-Jamendo (Bogdanov et al., 2019) for cross-dataset evaluation. Key detection uses AAM (Ostermann et al., 2023) and GiantSteps (Knees et al., 2015); genre tagging uses MagnaTagATune (Law et al., 2009) and GTZAN (Tzanetakis & Cook, 2002); and environmental sound recognition uses FSD50K (Fonseca et al., 2022) and ESC-50 (Piczak, 2015). The first three tasks are music-centered, while environmental sound recognition serves as a non-music control.

All datasets are mapped to fixed task-specific vocabularies: 20 instruments, 24 major/minor keys, 10 genres, and 12 environmental-sound categories. Multi-label tasks are evaluated with macro-F1, and key detection is evaluated with accuracy. All audio is represented as 10-second excerpts, using random excerpts during training and center excerpts during evaluation; shorter clips are zero-padded. For generation-based models, prompts specify the task, full label vocabulary, and constrained output format. We use greedy decoding and normalize generated text before mapping it to the canonical vocabulary with a task-specific parser; outputs with no valid label are counted as incorrect for single-label tasks, while multi-label outputs are scored as parsed label sets.

3. Experiment 1: Probing Existing AudioLMs

3.1. Goal and Method

We first ask where task information is accessible inside existing AudioLMs. To separate internal information from final text generation, we probe frozen representations from three models: MusiLingo (Deng et al., 2024), SALMONN (Tang et al., 2024), and Pengi (Deshmukh et al., 2023).

For each model, we extract encoder outputs, connector audio tokens, and LLM audio-token hidden states, using mean pooling in each case. We train a linear probe on each frozen representation using the task training split and evaluate it on the in-domain test split.

3.2. Results

Figure 2 shows that task information is often accessible beyond the encoder representation. Encoder-stage probes are frequently strongest, but the connector and LLM-audio stages do not show a uniform collapse: several trajectories remain nearly flat across stages, such as SALMONN on instrument, genre, and environmental sound recognition, and Pengi on environmental sound recognition.

The pattern is model- and task-dependent rather than a single monotonic trend. Some tasks lose probe accuracy after the encoder, while others remain stable across the connector and LLM-audio stages. Thus, poor closed-set recognition cannot be explained solely by a simple collapse of task information after the encoder. This motivates a controlled comparison between direct classification and LLM-based label generation.

4. Experiment 2: Controlled AudioLMs

4.1. Goal and Method

We compare two supervised ways of using the same frozen audio encoder for recognition: encoder-only classification and LLM-based label generation. The encoder-only baseline trains an MLP head on the encoder representation and predicts labels from logits. The LLM-generation model connects the same encoder to Qwen2.5-7B-Instruct (Qwen et al., 2025) through an MLP connector; the base LLM is frozen, and only the connector and LoRA adapters (Hu et al., 2022) are trained to generate labels in a fixed task-specific text format, parsed into canonical labels before scoring. Because the audio encoder is frozen in both routes, this comparison isolates the downstream recognition formulation after a fixed audio representation, rather than encoder adaptation.

We evaluate four encoders with complementary pretraining biases: CLAP (Elizalde et al., 2023), an audio-text contrastive encoder; BEATs (Chen et al., 2023), a general audio SSL encoder; MERT (Li et al., 2024), a music-focused SSL encoder; and PANNs (Kong et al., 2020), an AudioSet-supervised audio-tagging encoder. All models use the data splits and 10-second audio protocol defined in Section 2. LLM-generation models use LoRA rank 8, alpha 32, and dropout 0.1.

4.2. Results

Table 1 shows that the two downstream formulations favor different settings. LLM generation substantially improves in-domain instrument recognition on average, but

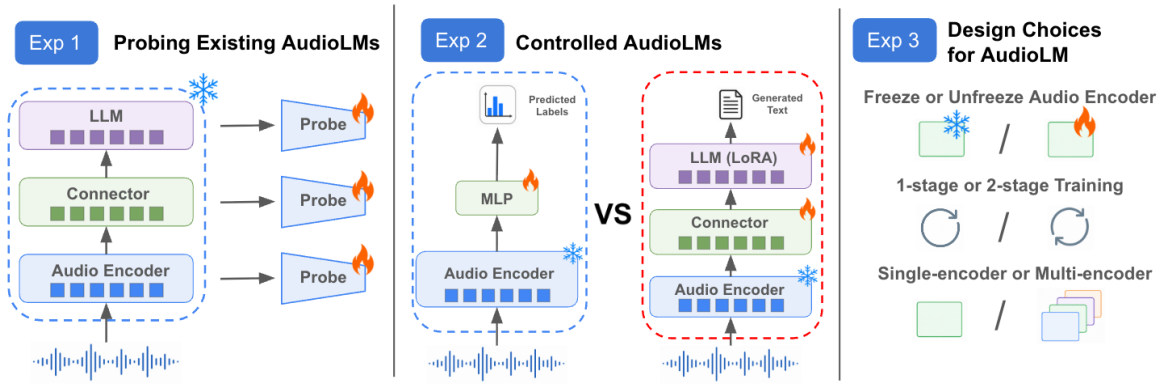

Figure 1. Overview of our three-part study. Experiment 1 probes existing AudioLMs at the audio encoder, connector, and LLM audio-token stages to test whether task-relevant information remains accessible. Experiment 2 compares encoder-only classification with AudioLM-style label generation under a controlled setting with the same frozen audio encoder and a shared LLM backbone. Experiment 3 evaluates encoder and connector design choices for joint AudioLM recognition: encoder tuning, caption-based connector pretraining, and single- versus multi-encoder aggregation.

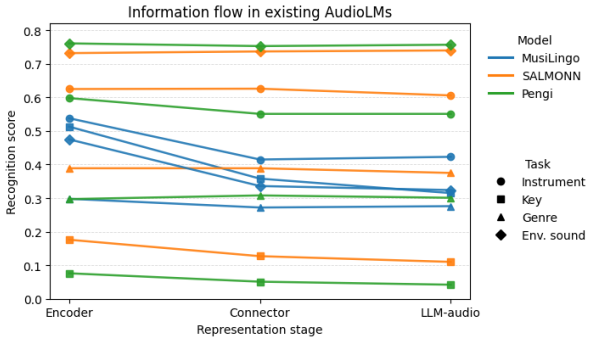

Figure 2. Linear probes on existing AudioLM representations.

Table 1. Controlled comparison between encoder-only classification and LLM-based label generation, averaged over four encoders. Δ = LLM − Enc.; W counts LLM wins. Multi-label tasks use macro-F1; key uses accuracy.

	In-domain			Cross-data		
Task	Enc. ↑	LLM ↑	Δ/W	Enc. ↑	LLM ↑	Δ/W
Instrument	.426	**.596**	+.169/4	**.235**	.149	-.086/0
Key	**.256**	.255	-.001/2	**.122**	.105	-.017/1
Genre	**.275**	.229	-.046/0	**.339**	.227	-.112/0
Env. sound	.641	**.694**	+.053/3	.779	**.871**	+.092/4

the direction reverses under cross-dataset transfer. Genre tagging consistently favors encoder-only classification, with negative deltas in both evaluation settings and no LLM-generation wins. Environmental sound recognition shows the opposite pattern, especially in cross-dataset evaluation, where LLM generation improves all four encoders. Key detection is mixed, with small average changes and no consistent advantage for either formulation.

Overall, using an LLM as a label-generation interface changes recognition behavior rather than acting as a uniformly stronger classifier. The effect depends on both task and dataset setting: it is clearest for environmental sound recognition, consistently negative for genre tagging, and mixed for instrument and key recognition. These trends may also reflect differences in label mappings and annotation conventions across datasets, which can affect generation-based evaluation.

Table 2. Effect of encoder tuning in joint four-task training. Δ is the change in joint score relative to the frozen setting for the same encoder. Joint score is the mean of the four task metrics.

	ID Δ ↑		Cross-data Δ ↑	
Encoder	Small-LR	Uniform	Small-LR	Uniform
BEATs	+0.042	-0.118	+0.002	-0.113
CLAP	+0.041	+0.057	+0.020	-0.006
MERT	+0.022	+0.081	+0.022	+0.044
PANNs	+0.018	+0.030	+0.020	+0.023

5. Experiment 3: Design Choices for Joint AudioLM Recognition

Experiment 2 shows that LLM-based generation is not uniformly better than direct classification and that its effect depends strongly on the task, dataset setting, and encoder. We therefore shift from asking whether the LLM interface helps by itself to asking which encoder and connector design choices improve joint AudioLM recognition. We evaluate three choices under joint four-task training: encoder tuning, caption-based connector pretraining, and multi-encoder aggregation.

5.1. Encoder Tuning

We compare three joint-training settings: frozen, where only the connector and LoRA parameters are trained; small-LR, where the encoder uses learning rate 10^{-5} and the connector and LoRA parameters use 10^{-4}; and uniform, where all trainable components use 10^{-4}.

Table 2 shows that small-LR tuning is the more reliable joint-training recipe: it improves the joint score for all four en-

Table 3. Effect of AudioCaps connector pretraining on joint cross-dataset performance. Joint score is the mean of the four task metrics.

Encoder	1-stage ↑ no pretrain	2-stage ↑ AudioCaps
BEATs	**0.360**	0.343
CLAP	**0.344**	0.337
MERT	0.286	**0.312**
PANNs	**0.327**	0.316

coders in-domain and is neutral or positive in cross-dataset evaluation. The task-level deltas are not uniformly positive, but most changes are small or favorable. Uniform tuning can give larger gains for selected encoder–task pairs, especially MERT on genre and CLAP on key detection, but it also strongly degrades BEATs. We therefore treat small-LR as the safer default for encoder tuning.

5.2. Caption-Based Connector Pretraining

We next test whether a caption-alignment stage improves joint recognition. The one-stage baseline trains directly on the four-task recognition mixture. The two-stage model first trains only the connector on AudioCaps (Kim et al., 2019), with the audio encoder and LLM frozen, and then uses the same joint recognition training as the baseline.

Table 3 shows that caption pretraining is not a reliable transfer recipe for this setting. It only improves MERT, but the other encoders are flat or lower after AudioCaps pretraining. Thus, AudioCaps-based connector pretraining is not a reliable substitute for task-specific recognition training in this setup.

5.3. Multi-Encoder Aggregation

The previous experiments show that no single encoder is uniformly best across tasks: different encoders show different task preferences, suggesting complementary pretraining biases. We therefore test whether joint recognition benefits from combining complementary encoders. BEATs, CLAP, MERT, and PANNs are run in parallel on the same waveform.

Because the encoders produce sequences with different feature dimensions and temporal rates, each encoder output is first projected into a shared feature space. Let $X_e \in \mathbb{R}^{T_e \times d_e}$ denote the output sequence from encoder e. We apply an encoder-specific projection:

$$H_e = X_e W_e, \qquad H_e \in \mathbb{R}^{T_e \times d}. \tag{1}$$

We compare two fusion methods. Concat fusion resamples all H_e to a common temporal rate, concatenates them along the feature dimension, and projects the result into audio tokens. Attention fusion instead uses learnable audio queries to attend over the projected token sequences from all encoders:

$$Z = \mathrm{CrossAttn}\left(Q, [H_{\mathrm{BEATs}}; H_{\mathrm{CLAP}}; H_{\mathrm{MERT}}; H_{\mathrm{PANNs}}]\right). \tag{2}$$

This lets each output audio token draw from any encoder stream without first forcing all encoders onto the same temporal grid.

Table 4. Frozen multi-encoder aggregation in joint recognition. Joint scores are the mean of the four task metrics.

Audio representation	ID ↑	Cross-data ↑
Best single encoder	0.488	0.360
Multi-encoder concat	0.485	0.350
Multi-encoder attention	**0.521**	**0.365**

Table 4 shows that multi-encoder aggregation helps only in the attention-fusion setting. Concat fusion slightly underperforms the best single-encoder baseline, indicating that simply exposing the model to more encoder features is not sufficient. Attention fusion is more effective, improving the frozen in-domain joint score from 0.488 to 0.521, while cross-dataset performance is largely unchanged. A separate small-LR attention-fusion run reaches 0.544 in-domain and 0.367 cross-dataset, the strongest in-domain joint result in our experiments.

6. Limitations

Our study is limited to one LLM backbone, one connector design, and deterministic closed-set prompting. Cross-dataset trends may partly reflect mapped label spaces and annotation conventions. Joint scores average heterogeneous task metrics, and multi-encoder models are not capacity-matched to single-encoder baselines; their gains should therefore be interpreted as practical improvements rather than isolated evidence for fusion alone.

7. Conclusion

We studied whether AudioLMs improve closed-set audio recognition or mainly change how audio evidence is decoded into fixed labels. Probing existing AudioLMs shows that task information often remains accessible beyond the encoder, weakening a simple information-loss explanation. Controlled comparisons show that LLM-based label generation is not uniformly stronger than direct encoder classification: it helps environmental sound recognition, hurts genre tagging, and is mixed for instrument and key recognition. Joint-training experiments further show that encoder and connector choices remain central: small-LR encoder tuning and attention-based multi-encoder fusion give the strongest results in our setup. Overall, closed-set AudioLM recognition depends not only on a language interface, but on how audio representations are adapted, combined, and mapped to labels.

Acknowledgments

This work is supported in part by funding from the Social Sciences and Humanities Research Council of Canada (SSHRC) under Grant No. 895-2022-1004.

References

Bogdanov, D., Won, M., Tovstogan, P., Porter, A., and Serra, X. The MTG-Jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning*, 2019.

Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., Che, W., Yu, X., and Wei, F. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 5178–5193, 2023.

Cooper, A., Kato, K., Shih, C.-H., Yamane, H., Vinken, K., Takemoto, K., Sunagawa, T., Yeh, H.-W., Yamanaka, J., Mason, I., et al. Rethinking VLMs and LLMs for image classification. *Scientific Reports*, 15(1):19692, 2025.

Deng, Z., Ma, Y., Liu, Y., Guo, R., Zhang, G., Chen, W., Huang, W., and Benetos, E. MusiLingo: Bridging music and text with pre-trained language models for music captioning and query response. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3643–3655, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.231.

Deshmukh, S., Elizalde, B., Singh, R., and Wang, H. Pengi: An audio language model for audio tasks. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Elizalde, B., Deshmukh, S., Al Ismail, M., and Wang, H. CLAP: Learning audio concepts from natural language supervision. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

Fonseca, E., Favory, X., Pons, J., Font, F., and Serra, X. FSD50K: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2022. doi: 10.1109/TASLP.2021.3133208.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Humphrey, E. J., Durand, S., and McFee, B. OpenMIC-2018: An open dataset for multiple instrument recognition. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*, pp. 438–444, 2018.

Kim, C. D., Kim, B., Lee, H., and Kim, G. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1011. URL https://aclanthology.org/N19-1011/.

Knees, P., Faraldo, A., Herrera, P., Vogl, R., Böck, S., Hörschläger, F., and Le Goff, M. Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections. In *Proceedings of the 16th International Society for Music Information Retrieval Conference*, pp. 364–370, 2015.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., and Plumbley, M. D. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. doi: 10.1109/TASLP.2020.3030497.

Law, E., West, K., Mandel, M. I., Bay, M., and Downie, J. S. Evaluation of algorithms using games: The case of music tagging. In *Proceedings of the 10th International Society for Music Information Retrieval Conference*, pp. 387–392, 2009.

Li, Y., Yuan, R., Zhang, G., Ma, Y., Chen, X., Yin, H., Xiao, C., Lin, C., Ragni, A., Benetos, E., Gyenge, N., Dannenberg, R., Liu, R., Chen, W., Xia, G., Shi, Y., Huang, W., Wang, Z., Guo, Y., and Fu, J. MERT: Acoustic music understanding model with large-scale self-supervised training. In *The Twelfth International Conference on Learning Representations*, 2024.

Ostermann, F., Vatolkin, I., and Ebeling, M. AAM: A dataset of artificial audio multitracks for diverse music information retrieval tasks. *EURASIP Journal on Audio, Speech, and Music Processing*, 2023(1):13, 2023. doi: 10.1186/s13636-023-00278-7.

Piczak, K. J. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 1015–1018, 2015. doi: 10.1145/2733373.2806390.

Qwen, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang,

J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 Technical Report, January 2025. URL https://arxiv.org/abs/2412.15115. arXiv:2412.15115 [cs.CL].

Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., and Zhang, C. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Tong, S., Brown, E., Wu, P., Woo, S., Middepogu, M., Akula, S. C., Yang, J., Yang, S., Iyer, A., Pan, X., et al. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

Tzanetakis, G. and Cook, P. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002. doi: 10.1109/TSA.2002.800560.