LabelBuddy: An Open Source Music and Audio Language Annotation Tagging Tool Using AI Assistance

Ioannis Prokopiou[1,2], Ioannis Sina[3], Agisilaos Kounelis[3], Pantelis Vikatos[2], Themos Stafylakis[1,4]

[1]Athens University of Economics and Business, [2]Orfium, [3]University of Patras, [4]Archimedes/Athena R.C.,

gian.prokopiou@aueb.gr, sinaioannis@gmail.com

agis@ceid.upatras.gr, pantelis@orfium.com, tstafylakis@aueb.gr

Abstract

The advancement of Machine learning (ML), Large Audio Language Models (LALMs), and autonomous AI agents in Music Information Retrieval (MIR) necessitates a shift from static tagging to rich, human-aligned representation learning. However, the scarcity of open-source infrastructure capable of capturing the subjective nuances of audio annotation remains a critical bottleneck. This paper introduces **LabelBuddy**, an open-source collaborative auto-tagging audio annotation tool designed to bridge the gap between human intent and machine understanding. Unlike static tools, it decouples the interface from inference via containerized backends, allowing users to plug in custom models for AI-assisted pre-annotation. We describe the system architecture, which supports multi-user consensus, containerized model isolation, and a roadmap for extending agents and LALMs. Code available at https://github.com/GiannisProkopiou/gsoc2022-Label-buddy.

1 Introduction

The quality of the datasets used to train AI models constitutes a significant factor in accuracy, reliability, and generalization (Picard et al., 2020). Despite standardization efforts through public repositories like Zenodo[1] and data lakes (Espinal et al., 2022), or benchmarks like DCASE[2] and MIREX[3], available resources are often task-specific. In addition, the creation of specialized datasets can be a challenging procedure, complicated, time-consuming, and is often laborious and supported by human review (Voigtlaender et al., 2021). Building on the tradition of cross-disciplinary impact, the intersection of Natural Language Processing (NLP) with music and audio presents a frontier where language and sound converge. Most audio content contains an inherent linguistic dimension, yet creating datasets to capture these multimodal synergies remains a laborious bottleneck.

The domain of Music Information Retrieval (MIR) is currently undergoing a transition from discriminative paradigms characterized by static tag classification to generative and reasoning-based approaches. The rise of Large Audio-Language Models (LALMs) such as Music Flamingo (Ghosh et al., 2025), Qwen-Audio (Chu et al., 2023), and Audio Flamingo 3 (Goel et al., 2025) has introduced new capabilities for "chain-of-thought" reasoning and conversational audio understanding. However, the efficacy of these models is heavily dependent on the quality of alignment with human intent. Recent surveys indicate that objective metrics often fail to capture aesthetic nuance, necessitating a pivot toward Reinforcement Learning from Human Feedback (RLHF) and rigorous subjective evaluation methodologies (Kader and Karmaker, 2025).

Current workflows are often fragmented, and users resort to disjointed workflows, separating data curation from the critical phase of manual subjective evaluation (e.g., MUSHRA, GoListen, or pairwise preference testing) (Schoeffler et al., 2018; Barry et al., 2021). Users use waveform-based tools for segmentation (Grover et al., 2020), separate platforms for text handling, and distinct software for subjective evaluation (e.g., WebMUSHRA (Schoeffler et al., 2018)). This separation hinders the development of efficient *Human-in-the-Loop* (HITL) pipelines, where the uncertainty of the model's output should drive data acquisition. Furthermore, the "crisis of metrics" in generative music, where objective scores like FAD fail to correlate with human perception (Kader and Karmaker, 2025) demands tools that can seamlessly transition from annotation to subjective preference ranking.

To address this, we present **LabelBuddy**, an open-source collaborative auto-tagging audio annotation tool equipped with AI assistance with:

[1]https://zenodo.org/

[2]https://dcase.community/

[3]https://www.music-ir.org/mirex

Proceedings of the 4th Workshop on NLP for Music and Audio (NLP4MusA 2026), pages 7–12

1. **Decoupled AI-Assistance:** An isolated containerized architecture injects model predictions via declarative YAML files. We provide pre-trained models like YOHO (Venkatesh et al., 2022), musicnn (Pons and Serra, 2019), PANNs (Kong et al., 2020), and LALMs like Music Flamingo (Ghosh et al., 2025) for AI-assisted pre-annotation tags to shift user effort from creation to verification, while approved labels can be used to fine-tune the models.

2. **Collaborative Consensus:** Native support for multi-user roles (manager, annotator, reviewer) to ensure ground-truth reliability.

3. **Hybrid Workflow Support:** An architecture designed to support both region-based tagging and subjective preference aggregation.

2 Related Work

This section reviews annotation platforms, HITL workflows for LALMs, and infrastructure for subjective evaluation and RLHF.

Annotation Platforms & Domain Specificity. The landscape of data curation significantly varies by modality. For text, tools like BRAT (Stenetorp et al., 2012), Paladin (Nghiem et al., 2021), and PubAnnotation (Kim and Wang, 2012) facilitate linguistic tagging, while specialized frameworks like CAT (Bartalesi Lenzi et al., 2012) and MDSWriter (Meyer et al., 2016) handle semantic efficiency and summarization respectively. Active learning strategies have been explored in text labeling (e.g., ActiveAnno (Wiechmann et al., 2021), APLenty (Nghiem and Ananiadou, 2018)). In the visual domain, tools like VIA (Dutta and Zisserman, 2019) and Annotation Web (Smistad et al., 2021) show the need for domain-specific interfaces.

In the audio domain, tools like Audino (Grover et al., 2020) and BAT (Meléndez Catalán et al., 2017) excel at temporal tasks like Sound Event Detection (SED) and salience, while library-based solutions like Aubio (Brossier et al., 2019) facilitate feature extraction. Others, like Gecko (Levy et al., 2019), focus on voice segmentation. However, these tools generally lack the decoupled AI architecture required for modern reasoning model use. Conversely, general-purpose HITL platforms like Label Studio (Tkachenko et al., 2020-2022) and Prodigy (Montani and Honnibal, 2018) offer robust backends but often restrict collaborative features such as reviewer roles and consensus metrics to paid enterprise tiers. Furthermore, they lack native support for musical structures (e.g., bars, beats) found in specialized audio tools (Cartwright et al., 2017). A comparison between LabelBuddy and other existing tools is shown in Table 1.

Table 1: Comparison of LabelBuddy with existing tools.

Tool	Audio Specific	Decoupled AI-Assist	Open Source	Collaboratory Consensus
Audino	✓	-	✓	-
BAT	✓	-	✓	-
Aubio	✓	-	✓	-
Gecko	✓	-	✓	-
Prodigy	-	✓	-	-
Label Studio (CE)	-	✓	✓	-
LabelBuddy	✓	✓	✓	✓

LALMs & HITL Workflows. The state-of-the-art has shifted from fixed-vocabulary auto-taggers to Large Audio-Language Models (LALMs) such as Audio Flamingo 3 (Goel et al., 2025) and Qwen-Audio (Chu et al., 2023). These utilize unified encoders for "chain-of-thought" reasoning. To align them, we adopt a "Single-Iteration" HITL approach. Recent studies in video annotation (Gutiérrez et al., 2025) demonstrate that simple model-assisted pre-annotation reduces time-on-task without degrading quality, a philosophy we extend to audio similarly to NEAL (Gibbons et al., 2023).

Subjective Evaluation & RLHF. A critical bottleneck in generative music is the "crisis of metrics," where scores like Fréchet Audio Distance (FAD) fail to correlate with human perception (Kader and Karmaker, 2025; Gui et al., 2023). Consequently, the field is pivoting towards RLHF (Cideron et al., 2024; Liu et al., 2025). Currently, evaluation is decoupled from annotation, relying on standalone tools like WebMUSHRA (Schoeffler et al., 2018). LabelBuddy aims to unify this by integrating pairwise preference aggregation methods, such as Bayesian Bradley–Terry (BBQ) (Aczel et al., 2025).

3 System Architecture

LabelBuddy addresses the "coupling problem" in annotation tools where interfaces are hard-coded to specific model backends via a modular, containerized architecture. As illustrated in Figure 1, the system decouples the lightweight user interaction layer (Django) from the compute-intensive inference layer (Docker).

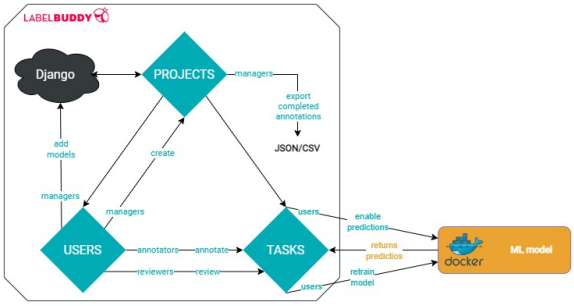

Figure 1: **System Architecture Overview**: The architecture decouples the Django web server from Dockerized ML inference.

3.1 Backend & Data Model

The core application is built on Django, utilizing a relational database to manage the three primary entities: *Projects*, *Users*, and *Tasks*.

- **RBAC & Privacy:** To prevent data leakage, the system implements Role-Based Access Control (RBAC). Managers have full oversight, while *Annotators* and *Reviewers* are restricted to their assigned task queues.
- **Data Serialization:** Annotations are stored as JSON objects that contain temporal boundaries and ontology tags. Managers can export consensus data to CSV/JSON formats for direct integration with ML training pipelines.

3.2 Containerized Inference Engine

Managers define models via a YAML configuration file, specifying the Docker image, input/output schema, and resource constraints. When AI assistance is requested, the backend communicates with the model container via a RESTful Flask API. This design ensures Sandboxing (models run in isolated environments) and Scalability (inference can be deployed on remote cloud nodes like AWS/Azure).

4 Workflow & Interface

The platform supports a comprehensive "Human-in-the-Loop" (HITL) lifecycle.

4.1 Project Setup & Task Management

The workflow begins in the **Dashboard**, where managers create projects and assign user roles.

- **Model Integration:** In the **Model Page**, managers upload YAML configuration files to attach inference containers. This interface exposes advanced controls: monitoring training loss/accuracy, downloading weight files, and triggering fine-tuning jobs (specifying epochs and learning rates) using the project's validated data.
- **Task Ingestion:** Managers upload audio (WAV/MP3) via the **Project Page**, which can be distributed to annotators via a shared pool or disjoint assignment strategies.

4.2 The Annotation Loop

The core labeling workflow utilizes `wavesurfer.js` for responsive waveform visualization.

1. **AI-Assisted Pre-Annotation:** Annotators trigger "On-Demand Prediction," which serializes the audio to the active Docker container. The system renders the returned predictions as editable regions (Fig. 2), shifting the human task from *creation* to *verification*.
2. **Review & Consensus:** Completed tasks enter the **Review Interface**, where reviewers can play back specific regions and approve or reject annotations with feedback. This Quality Assurance (QA) loop is essential for creating high-fidelity datasets for generative alignment.

5 Case Study: NLP Music Tagging

To demonstrate LabelBuddy's utility, we present a reference workflow for creating a Music Captioning Dataset, a task requiring the alignment of audio signals with rich natural language descriptions.

Model Integration The project manager defines a Docker container that wraps a multimodal model, such as a Music Flamingo checkpoint. The YAML configuration maps the model's text output to a LabelBuddy `Annotation` region:

```
image: "my-repo/music-flamingo:v1"
input_schema: { "audio": "wav" }
output_schema:
  - { "type": "text", "label": "Caption" }
resources: { "gpu": "true" }
```

The "Human-Verify" Loop. Annotators are presented with a queue of raw audio tracks. Instead of writing descriptions from scratch (which is cognitively demanding), they trigger the **Pre-Annotate** function. The backend container processes the audio and returns a candidate caption: *"A lo-fi hip-hop track with a slow tempo and vinyl crackle."*

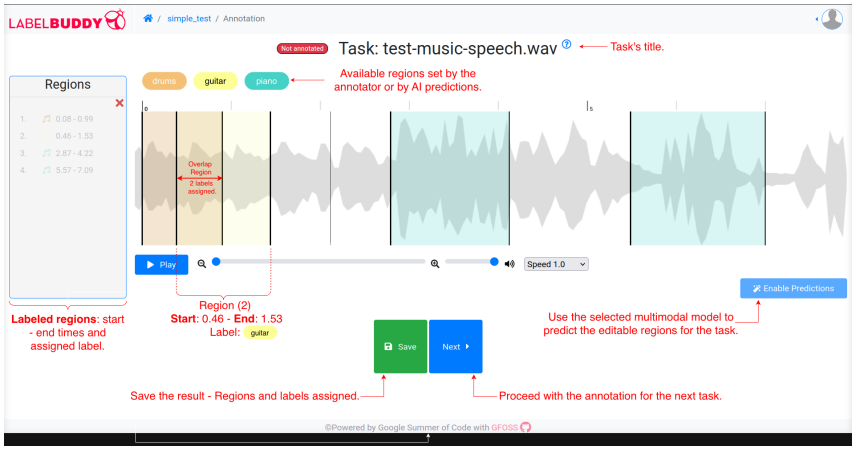

Figure 2: The **annotation interface** displaying AI-generated predictions as editable waveform regions.

Correction & Consensus. The annotator corrects specific hallucinations (e.g., changing "vinyl crackle" to "rain sounds") and adjusts timestamp boundaries. If multiple annotators process the same track, the Reviewer Interface highlights semantic disagreements in the captions, allowing for the creation of robust, consensus-based ground truth.

Multimodal Export. The finalized dataset is exported as a JSONL or CSV file containing aligned `(audio_path, text_caption)` pairs, ready for immediate use in fine-tuning downstream audio-to-text generation models.

6 Discussion & Future Roadmap

The current release of LabelBuddy solves the immediate infrastructure challenge: decoupling the annotation frontend from rapidly evolving model backends. As the field transitions towards LALMs and autonomous agents, our future roadmap is aligned with bridging the gap between human intent and machine representation:

Agentic Reasoning. While traditional active learning relies on uncertainty sampling, the rise of LALMs requires a shift toward *conversational* assistance. We are extending the backend API to support more reasoning-capable models such as Qwen-Audio (Chu et al., 2023). Future versions will allow annotators to query the model and receive "Chain-of-Thought" justifications, transforming the workflow from simple tag verification to collaborative reasoning. This aligns with recent findings that interactive reasoning reduces hallucination in complex annotation tasks.

Integrated Subjective Evaluation (RLHF). Acknowledging the "crisis of metrics" where FAD scores fail to capture aesthetic quality (Kader and Karmaker, 2025), LabelBuddy aims to evolve into a workbench for RLHF. We aim to implement a native "Pairwise Preference" interface (Dataset A vs. Dataset B) directly in the review loop. To handle noisy human raters, the backend will integrate Bayesian Bradley–Terry (BBQ) models (Aczel et al., 2025), providing robust preference aggregation to align generative models (Cideron et al., 2024).

Enhancing Perceptual Validity. To counteract the tendency of models to rely on text priors rather than audio content a flaw highlighted by the RUListening benchmark (Zang et al., 2025), we plan to introduce timestamp-required QA templates. These will force both models and human annotators to ground every semantic claim in specific spectral regions, ensuring that future datasets drive genuine auditory perception rather than text-only reasoning.

Evaluation Plan. To validate utility, we propose a pilot study on DCASE 2024 data measuring: (a) time reduction vs. *de novo* labeling, (b) inter-annotator agreement (Fleiss' Kappa), and (c) downstream PSDS gains for baseline SED models trained on LabelBuddy-curated data.

7 Conclusion

LabelBuddy serves as critical infrastructure for exploring the multimodal synergies between language and audio. By decoupling the interface from inference, it empowers the community to curate the rich, linguistically-grounded datasets required for modern NLP-driven music understanding. Whether for standard tagging or the emerging demands of RLHF, LabelBuddy offers an open, scalable workbench to deepen the connection between human perception and machine representation on audio.

8 Ethics Statement

The development of AI-assisted annotation tools raises concerns regarding labor displacement and bias. LabelBuddy is designed to augment, not replace, human expertise, keeping the human in the loop for critical judgments. Furthermore, by facilitating the creation of open datasets, we aim to democratize access to high-quality training data, countering the centralization of resources in large tech corporations. We ensure that all integrated models are used in compliance with their research licenses.

Acknowledgments

This work was supported by Google Summer of Code (GSoC) and the Open Technologies Alliance (GFOSS) and was funded by the European Union's Horizon Europe research and innovation programme under the AIXPERT project (Grant Agreement No. 101214389), which aims to develop an agentic, multi-layered, GenAI-powered framework for creating explainable, accountable and transparent AI systems.

References

Till Aczel, Lucas Theis, and Wattenhofer Roger. 2025. Efficient bayesian inference from noisy pairwise comparisons. *arXiv preprint arXiv:2510.09333*.

Dan Barry, Qijian Zhang, Pheobe Wenyi Sun, and Andrew Hines. 2021. Go listen: an end-to-end online listening test platform. *Journal of Open Research Software*, 9(1).

Valentina Bartalesi Lenzi, Giovanni Moretti, and Rachele Sprugnoli. 2012. Cat: the celct annotation tool. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 333–338. LREC.

Paul Brossier, Tintamar, Eduard Müller, Nils Philippsen, Tres Seaver, Hannes Fritz, cyclopsian, Sam Alexander, Jon Williams, James Cowgill, and Ancor Cruz. 2019. aubio/aubio: 0.4.9.

Mark Cartwright, Ayanna Seals, Justin Salamon, Alex Williams, Stefanie Mikloska, Duncan MacConnell, Edith Law, Juan P Bello, and Oded Nov. 2017. Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–21.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Geoffrey Cideron, Sertan Girgin, Mauro Verzetti, Damien Vincent, Matej Kastelic, Zalán Borsos, Brian McWilliams, Victor Ungureanu, Olivier Bachem, Olivier Pietquin, et al. 2024. Musicrl: Aligning music generation to human preferences. *arXiv preprint arXiv:2402.04229*.

Abhishek Dutta and Andrew Zisserman. 2019. The via annotation software for images, audio and video. In *Proceedings of the 27th ACM international conference on multimedia*, pages 2276–2279.

Xavier Espinal, Maria Giuffrida, Marieke Willems, and Rita Meneses. 2022. Bringing big science experiment data to the researchers' fingertips.

Sreyan Ghosh, Arushi Goel, Lasha Koroshinadze, Sanggil Lee, Zhifeng Kong, Joao Felipe Santos, Ramani Duraiswami, Dinesh Manocha, Wei Ping, Mohammad Shoeybi, et al. 2025. Music flamingo: Scaling music understanding in audio language models. *arXiv preprint arXiv:2511.10289*.

Anthony Gibbons, Ian Donohue, Courtney Gorman, Emma King, and Andrew Parnell. 2023. Neal: an open-source tool for audio annotation. *PeerJ*, 11:e15913.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*.

Manraj Singh Grover, Pakhi Bamdev, Yaman Kumar, Mika Hama, and Rajiv Ratn Shah. 2020. audino: A modern annotation tool for audio and speech. *arXiv preprint arXiv:2006.05236*.

Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. 2023. Adapting frechet audio distance for generative music evaluation. *arXiv preprint arXiv:2311.01616*.

Juan Gutiérrez, Ángel Mora, Pablo Regodón, Silvia Rodriguez, and José Luis Blanco. 2025. Ai-boosted video annotation: Assessing the process enhancement. *arXiv preprint arXiv:2510.21798*.

Faria Binte Kader and Santu Karmaker. 2025. A survey on evaluation metrics for music generation. *arXiv preprint arXiv:2509.00051*.

Jin-Dong Kim and Yue Wang. 2012. Pubannotation-a persistent and sharable corpus and annotation repository. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 202–205.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns:

Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

Golan Levy, Raquel Sitman, Ido Amir, Eduard Golshtein, Ran Mochary, Eilon Reshef, Roi Reichart, and Omri Allouche. 2019. Gecko-a tool for effective annotation of human conversations. In *INTERSPEECH*, pages 3677–3678.

Cheng Liu, Hui Wang, Jinghua Zhao, Shiwan Zhao, Hui Bu, Xin Xu, Jiaming Zhou, Haoqin Sun, and Yong Qin. 2025. Musiceval: A generative music dataset with expert ratings for automatic text-to-music evaluation. *arXiv preprint arXiv:2501.10811*.

Blai Meléndez Catalán, Emilio Molina, and Emilia Gómez Gutiérrez. 2017. Bat: An open-source, web-based audio events annotation tool.

Christian M Meyer, Darina Benikova, Margot Mieskes, and Iryna Gurevych. 2016. Mdswriter: annotation tool for creating high-quality multi-document summarization corpora. In *Proceedings of ACL-2016 System Demonstrations*, pages 97–102.

Ines Montani and Matthew Honnibal. 2018. Prodigy: A new annotation tool for radically efficient machine teaching. *Artificial Intelligence to appear*.

Minh-Quoc Nghiem and Sophia Ananiadou. 2018. Aplenty: annotation tool for creating high-quality datasets using active and proactive learning. In *EMNLP (Demonstration)*, pages 108–113.

Minh-Quoc Nghiem, Paul Baylis, and Sophia Ananiadou. 2021. Paladin: an annotation tool based on active and proactive learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 238–243.

Sylvaine Picard, Camille Chapdelaine, Cyril Cappi, Laurent Gardes, Eric Jenn, Baptiste Lefevre, and Thomas Soumarmon. 2020. Ensuring dataset quality for machine learning certification. In *2020 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, pages 275–282. IEEE.

Jordi Pons and Xavier Serra. 2019. musicnn: Pre-trained convolutional neural networks for music audio tagging. *arXiv preprint arXiv:1909.06654*.

Michael Schoeffler, Sarah Bartoschek, Fabian-Robert Stöter, Marlene Roess, Susanne Westphal, Bernd Edler, and Jürgen Herre. 2018. webmushra—a comprehensive framework for web-based listening tests. *Journal of Open Research Software*, 6(1).

Erik Smistad, Andreas Østvik, and Lasse Løvstakken. 2021. Annotation web-an open-source web-based annotation tool for ultrasound images. In *2021 IEEE International Ultrasonics Symposium (IUS)*, pages 1–4. IEEE.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Satvik Venkatesh, David Moffat, and Eduardo Reck Miranda. 2022. You only hear once: a yolo-like algorithm for audio segmentation and sound event detection. *Applied Sciences*, 12(7):3293.

Paul Voigtlaender, Lishu Luo, Chun Yuan, Yong Jiang, and Bastian Leibe. 2021. Reducing the annotation effort for video object segmentation datasets. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3060–3069.

Max Wiechmann, Seid Muhie Yimam, and Chris Biemann. 2021. Activeanno: General-purpose document-level annotation tool with active learning integration. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 99–105.

Yongyi Zang, Sean O'Brien, Taylor Berg-Kirkpatrick, Julian McAuley, and Zachary Novack. 2025. Are you really listening? boosting perceptual awareness in music-qa benchmarks. *arXiv preprint arXiv:2504.00369*.