From Generation to Attribution: Music AI Agent Architectures for the Post-Streaming Era

Wonil Kim[1*], **Hyeongseok Wi**[1*], **Seungsoon Park**[1*], **Taejun Kim**[1*], **Sangeun Keum**[1*], **Keunhyoung Kim**[1*], **Taewan Kim**[1*], **Jongmin Jung**[1], **Taehyoung Kim**[1], **Gaetan Guerrero**[1], **Mael Le Goff**[1], **Julie Po**[1], **Dongjoo Moon**[1], **Juhan Nam**[1,2], **Jongpil Lee**[1*]

[1]MixAudio by Neutune, [2]KAIST

Abstract

Generative AI is reshaping music creation, but its rapid growth exposes structural gaps in attribution, rights management, and economic models. Unlike past media shifts—from live performance to recordings, downloads, and streaming—AI transforms the entire lifecycle of music, collapsing boundaries between creation, distribution, and monetization. However, existing streaming systems, with opaque and concentrated royalty flows, are ill-equipped to handle the scale and complexity of AI-driven production.

We propose a *content-based Music AI Agent* architecture that embeds attribution directly into the creative workflow through block-level retrieval and agentic orchestration. Designed for iterative, session-based interaction, the system organizes music into granular components (*Blocks*) stored in *BlockDB*; each use triggers an *Attribution Layer* event for transparent provenance and real-time settlement.

This framework reframes AI from a generative tool into infrastructure for a *Fair AI Media Platform*. By enabling fine-grained attribution, equitable compensation, and participatory engagement, it points toward a *post-streaming paradigm* where music functions not as a static catalog but as a collaborative and adaptive ecosystem. Session example demo: `https://aimusicagent.github.io/`.

1 Introduction

Generative AI is emerging as a transformative force in the music industry. Advances in cross-modal generation, instrumental synthesis, and voice modeling have enabled music to be created instantly, often in the form of single-shot outputs [1–3]. While this single-shot mode demonstrates technical progress, it limits the depth of creative interaction and obscures the role of attribution. Recent commercial systems (e.g., text-to-music platforms) typically deliver end-to-end songs without offering iterative refinement or mechanisms for crediting contributors. This gap highlights a structural misalignment between technological capability and the requirements of sustainable creative ecosystems. To move beyond this, there is a growing need for systems that treat composition not as a one-time act but as an iterative and participatory process.

As illustrated in Figure 1, the history of the music industry has been shaped by successive shifts in media formats, each requiring new legal and economic frameworks. From copyright in live performance to mechanical rights in recorded music, neighboring rights in the CD era, and royalty accounting in streaming, every transition has redefined how rights are attributed and revenues distributed [4, 5]. Unlike earlier shifts, which expanded rights around tangible media, generative AI collapses creation, distribution, and monetization into a single process. This convergence creates

[*]Equal contribution. Corresponding author: `ceo@mix.audio`

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Music.

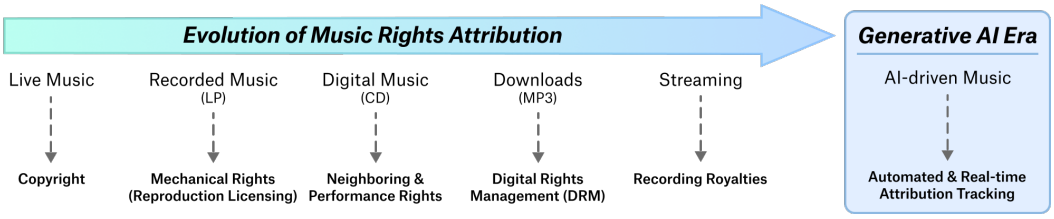

Figure 1: Music rights across media shifts into the AI era.

demands for real-time, fine-grained attribution and accounting frameworks that existing models cannot yet provide [6].

This gap collides with the structural limitations of today's streaming economy. Subscription-based royalty models remain opaque and concentrated among a few artists, while metadata fragmentation and fraudulent practices exacerbate the "black box" of unmatched royalties [5, 7, 8]. As AI dramatically scales both the volume and diversity of music, these inefficiencies risk being magnified: absent attribution-aware infrastructure, independent artists and cultural communities risk seeing their contributions absorbed into opaque accounting with limited recognition or compensation. In other words, AI could enlarge the very inequities it is often expected to solve.

To overcome these limitations, we argue that AI systems must extend beyond generation to encompass attribution, participation, and monetization. Embedding rights tracking not only into the creative process but also into modes of consumption is essential for building sustainable platforms. This trajectory points toward a post-streaming paradigm: an adaptive, participatory, and collaborative ecosystem where music is no longer a static catalog but a living process of creation and engagement.

In this paper, we propose a Music AI Agent architecture that operationalizes this vision. Built on content-based Retrieval-Augmented Generation (RAG), generative models, and agent-based orchestration, the system embeds attribution directly into the creative workflow. By recording *Block-level* usage events—where *Blocks* denote granular musical components such as stems or structural segments—the system directly links creation with transparent provenance. Combined with real-time royalty settlement, this lays the foundation for what can be described as a *Fair AI Media Platform* that aligns generative capabilities with transparent attribution and equitable economic models for the future of music.

2 Background

In parallel with advances in LLMs, RAG, and AI Agents in the language domain, we explore how these paradigms can be extended to media content-based AI Agent systems, with a particular focus on references in the domain of music.

2.1 Generative Model

As LLMs serve as foundational models in language, their counterparts in music audio are transformer- [2, 9, 10], diffusion- [3, 11–14], and flow-matching–based [15] generative models, enabling full-song [11], stem [16–19], and singing voice [20–24] generation or conversion [25]. Services such as Suno [26], Udio [27], and Riffusion [28] offer full-song generation but lack transparency in training data and revenue sharing [29]. In response, ElevenLabs [30] has partnered with independent music distributors like Kobalt [31] and Merlin [32], and Beatoven.ai [33] with providers such as Musical AI, though current revenue-sharing models remain rudimentary (e.g., simple 1/N splits) [34]. More recently, Splice has introduced sample generation and transformation tools [35], and Moises supports context-aware stem generation—together [36] signaling both the diversification of AI music services and the need for richer, more interactive user experiences.

2.2 Content-based RAG

Just as LLMs have evolved into RAG systems with Memory DBs to support source attribution and richer user experiences [37], the analogue for media content lies in generative models grounded in a database of component blocks of media [38–41]. In these multimodal RAG systems, retrieval is not limited to text but extends across heterogeneous modalities such as visual content and motion data, where textual queries can be jointly matched with image or motion embeddings to enhance specificity

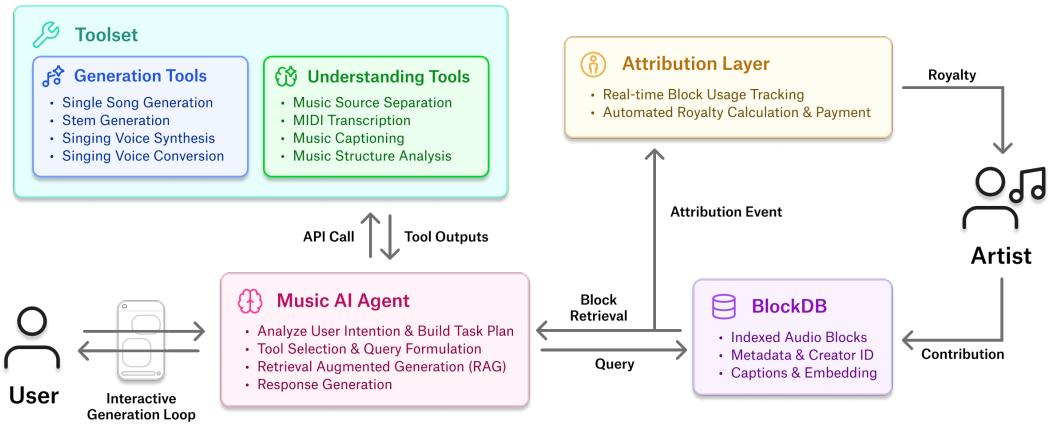

Figure 2: The architecture of the proposed Music AI agent designed to ensure fair artist compensation through an automated attribution and royalty system.

of the media content. In music, this translates to a music sample database from which elements are retrieved via text queries and then recombined to generate new content. Such an approach enables explicit reference tracking of samples, facilitating attribution and revenue sharing. Structurally, the mapping aligns with RAG: indexing and parsing correspond to Music Information Retrieval (MIR) [42], while retrieval, language–audio models [43–46], and generative models form the pipeline [19, 47–50].

2.3 AI Agent Architecture containing Media Content

Extending content-based RAG with an agent-based orchestration layer and toolsets yields a highly organic user experience in music creation and transformation. Early explorations such as Music AI Agents [51] integrated diverse MIR techniques into toolsets, enabling users not only to generate music but also to isolate and manipulate its components. However, these systems remain algorithmically toolset-driven [52]. If evolved into agent architectures backed by media-content memory databases—where samples or artist tracks are decomposed into granular musical elements—such systems could further support attribution and enable revenue-sharing mechanisms proportional to original source contributions.

3 Content-based Music AI Agent for Integrated Attribution

To address the core challenge of attribution, we introduce an architecture that operationalizes the paradigms of content-based RAG and AI agents. Our system orchestrates a suite of generative tools grounded in a database of musical components, thereby making attribution an intrinsic part of the creative act rather than a post-hoc process. As detailed in Figure 2, this architecture is composed of three pillars: a **BlockDB** of artist-contributed media, an interactive **Music AI Agent** with a versatile **toolset**, and a real-time **Attribution Layer**.

Creator Contribution Workflow The process begins with artists submitting original music to the platform, where it is analyzed, separated into *Blocks*, and indexed in BlockDB with metadata including its musical properties, descriptive embeddings, and a record of its original creator. Blocks are created either by decomposing artist tracks—supporting attribution and distribution—or by incorporating stems and samples from producers, which complement the database to address the existing copyright frameworks not covered by decomposition process alone. This process builds the library of attributable musical assets that powers the ecosystem.

Interactive Generation and Real-time Attribution The creation process unfolds as a **session-based, interactive loop**, where users collaborate with the Music AI Agent to build a track layer by layer. In each turn of this creative dialogue, the agent executes a RAG workflow by retrieving relevant Blocks from BlockDB in response to user prompts. These Blocks then serve as context or creative conditions for the diverse generative tools in the agent's toolset. Crucially, each time a Block is used to generate a new musical layer, it triggers a real-time **attribution event**, linking the new output to its original creators. The Attribution Layer aggregates these events throughout the session and

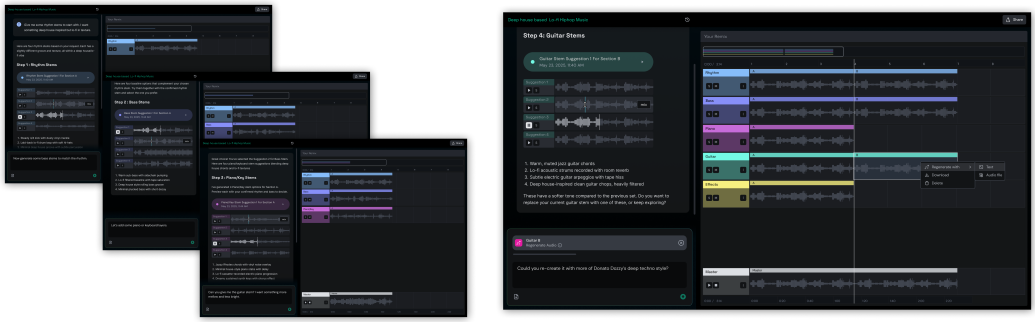

Figure 3: The user interface for a Music AI Agent, showing how a song is built through session-based, interactive layering of musical stems (left), and how a user can then select a specific track to request further modifications (right).

allocates royalties to the original artists under transparent rules, completing the cycle from creation to compensation. By embedding attribution within this iterative RAG pipeline, the architecture ensures that compensation reflects a work's actual granular use, rather than approximations based on market share. This design provides the foundation for a more equitable AI-driven music economy, which is explored further in the following section.

4 Discussion and Conclusion

The introduction of attribution-centered Music AI Agents represents more than a technical innovation. It provides a structural alternative to the existing streaming paradigm and signals the emergence of next-generation AI-driven media platforms.

Transformation of Creative Practice If such an agentic workflow is structured as iterative, session-based stem generation as illustrated in Figure 3, music creation is reframed as a participatory process that lowers barriers to entry and diversifies engagement. Novices can explore through conversational interaction (e.g., **vibe producing** or **remixing**), while professionals can pursue more refined creations with precise control. In this way, music is redefined not as a static product but as a living, adaptive medium of expression [10, 11].

AI Attribution By embedding attribution directly into the creative process, the system creates verifiable and auditable records at the level of components. This replaces the aggregate and opaque accounting of traditional streaming with transparent, traceable flows of credit and compensation. Every creative action becomes a documented contribution, strengthening accountability among artists, rights holders, and users, and positioning AI as a legally reliable partner in the creative cycle [53].

Economic Models and Fan Participation Granular contribution records enable economic models that move beyond blanket or pool-based subscription royalties. Instead, a usage-based micropayment mechanism could be a solution to current streaming fraud by directly tying rewards to creative contributions. Furthermore, fans evolve from passive listeners to active stakeholders who remix, co-create, and directly support artists. This dynamic expands the scope of the "superfan economy" into a broader ecosystem of participatory patronage [54].

Infrastructure and Copyright Integration Event logging through the Attribution Layer establishes infrastructure compatible with global copyright registries such as ISRC, ISWC, and IPI, positioning Music AI Agents as complementary extensions of existing rights management systems that connect creation, AI generation, and distribution. By managing the music lifecycle through transparent contribution records and automated settlements, Music AI Agents help form the foundation of a Fair AI Media Platform and support a more sustainable, interoperable, and collaborative music ecosystem [55].

Post-Streaming Outlook Collectively, these mechanisms signal a shift from centralized streaming platforms toward agentic AI ecosystems where creation, consumption, and settlement converge in real time. In this post-streaming paradigm, AI evolves from a generative tool into a trusted mediator that governs attribution, evaluation, and economic interaction under transparent rules [56, 57].

Acknowledgements

This work was made possible through the support of the following institutions and collaborators.

Institutions: Korea Creative Content Agency (KOCCA), National Gugak Center, National Information Society Agency (NIA), Korea Culture Information Service Agency (KCISA), Artificial Intelligence Industry Cluster Agency (AICA), and Electronics and Telecommunications Research Institute (ETRI).

Team: Theerasak Charoenchob, Teeratep Weerapang, Azamat Atabayev, Abboskhon Sobirov, Nursultan Soodonbekov, Hyeyeon Park

Advisory: Yoon Kim, Woojin Cha

Lead Advisor: Soo-man Lee

References

[1] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever. Jukebox: A generative model for music. In: *arXiv preprint arXiv:2005.00341*. 2020.

[2] J. Copet, A. Défossez, G. Synnaeve, and Y. Adi. Simple and Controllable Music Generation. In: *Proceedings of the 24th International Society for Music Information Retrieval Conference (ISMIR)*. 2023.

[3] H.-W. Liu, G. Mittal, Y. Ren, and Z. Duan. MusicLDM: Enhancing Novelty in Text-to-Music Generation Using Beat-Synchronous Guidance. In: *Proceedings of the 12th International Conference on Learning Representations (ICLR)*. 2024.

[4] World Intellectual Property Organization (WIPO). *The Future of Copyright in the Digital Era*. `https://www.wipo.int/copyright/en/`. Accessed: 2025-08-29. 2021.

[5] IFPI. *Global Music Report 2023*. `https://www.ifpi.org`. Accessed: 2025-08-29. 2023.

[6] D. J. Gervais. Artificial Intelligence and Copyright Law. In: *Houston Law Review* **60** (2023), pp. 321–370.

[7] L. Marshall. The Problem with Music Streaming. In: *International Journal of Cultural Studies*. Vol. 18. 3. SAGE, 2015, pp. 270–285.

[8] A. Herstand. *How Streaming Services Pay Musicians*. `https://aristake.com/post/how-streaming-services-pay-musicians`. Accessed: 2025-08-29. 2020.

[9] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, I. Simon, C. Hawthorne, N. Shazeer, A. M. Dai, M. Hoffman, M. Dinculescu, and D. Eck. Music Transformer: Generating Music with Long-Term Structure. In: *International Conference on Learning Representations (ICLR)*. 2019.

[10] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, M. Sharifi, N. Zeghidour, and C. Frank. *MusicLM: Generating Music From Text*. 2023. arXiv: `2301.11325 [cs.SD]`.

[11] F. Schneider, K. Koutini, and G. Widmer. Moûsai: Text-to-Music Generation with Long-Context Latent Diffusion. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. 2024.

[12] Z. Evans, C. Carr, J. Taylor, S. H. Hawley, and J. Pons. Fast timing-conditioned latent audio diffusion. In: *Proceedings of the 41st International Conference on Machine Learning*. ICML'24. Vienna, Austria: JMLR.org, 2024.

[13] Z. Evans, J. D. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons. Stable audio open. In: *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2025, pp. 1–5.

[14] B. Evans, J. Engel, S. Ginosar, A. Roberts, et al. Long-form Music Generation with Latent Diffusion. In: *arXiv preprint arXiv:2404.10301* (2024).

[15] K. Prajwal, A. Maiti, A. Singh, A. Shukla, S. Sharma, and S. Ganapathy. MusicFlow: Text-to-Music via Flow Matching. In: *arXiv preprint arXiv:2410.20478* (2024).

[16] J. D. Parker, J. Spijkervet, K. Kosta, F. Yesiler, B. Kuznetsov, J.-C. Wang, M. Avent, J. Chen, and D. Le. Stemgen: A music generation model that listens. In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2024, pp. 1116–1120.

[17] S. Rouard, R. San Roman, Y. Adi, and A. Roebel. MusicGen-Stem: Multi-stem music generation and edition through autoregressive modeling. In: *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2025, pp. 1–5.

[18] Z. Kong, M. Liu, N. Wu, W. Lu, Y. Fu, Y. Li, W. Wang, and B. Yu. Multi-Track MusicLDM: Towards Versatile Music Generation with Latent Diffusion Model. In: *arXiv preprint arXiv:2405.12967* (2024).

[19] G. Mariani, I. Tallini, E. Postolache, M. Mancusi, L. Cosmo, and E. Rodolà. Multi-Source Diffusion Models for Music Source Separation and Generation. In: *International Conference on Learning Representations (ICLR)*. 2024.

[20] R. Huang, C. Cui, F. Chen, Y. Ren, J. Liu, Z. Zhao, B. Huai, and Z. Wang. SingGAN: Generative Adversarial Network For High-Fidelity Singing Voice Generation. In: *Proceedings of the 30th ACM International Conference on Multimedia*. MM '22. Lisboa, Portugal: Association for Computing Machinery, 2022, pp. 2525–2535.

[21] Y. Zhang, J. Cong, H. Xue, L. Xie, P. Zhu, and M. Bi. VISinger: Variational Inference with Adversarial Learning for End-to-End Singing Voice Synthesis. In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, pp. 7237–7241.

[22] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao. DiffSinger: Singing Voice Synthesis via Shallow Diffusion Mechanism. In: *Proceedings of the AAAI Conference on Artificial Intelligence* **36**.10 (June 2022), pp. 11020–11028.

[23] B. Zhuang, R. Li, Z. Wang, X. Chen, D. Geng, H. Wang, J. Zhang, and J. Shi. SPA-SVC: Self-supervised Pitch Augmentation for Singing Voice Conversion. In: *Interspeech*. 2024.

[24] X. Zhang, Y. Huang, H. Wu, Y. Ju, Z. Wu, and H. Meng. Leveraging Diverse Semantic-based Audio Pretrained Models for Singing Voice Conversion. In: *2024 IEEE Spoken Language Technology Workshop (SLT)*. 2024.

[25] Y. Bai, H. Chen, J. Chen, Z. Chen, Y. Deng, X. Dong, L. Hantrakul, W. Hao, Q. Huang, Z. Huang, et al. Seed-music: A unified framework for high quality and controlled music generation. In: *arXiv preprint arXiv:2409.09214* (2024).

[26] *Suno AI*. `https://www.suno.ai/`. Accessed: 2025-08-27.

[27] *Udio AI*. `https://www.udio.com/`. Accessed: 2025-08-27.

[28] S. Forsgren and H. Martiros. *Riffusion: Real-time Music Generation with Stable Diffusion*. `https://www.riffusion.com/`. Accessed: 2025-08-27. 2022.

[29] Music Business Worldwide. *Major record companies sue AI music generators Suno, Udio for 'mass infringement' of copyright*. `https://www.musicbusinessworldwide.com/major-record-companies-sue-ai-music-generators-suno-udio-for-mass-infringement-of-copyright/`. Accessed: 2025-08-29. 2024.

[30] *ElevenLabs*. `https://elevenlabs.io/`. Accessed: 2025-08-27.

[31] *Kobalt Music Group*. `https://www.kobaltmusic.com/`. Accessed: 2025-08-27.

[32] *Merlin Network: Independent Music Licensing*. `https://www.merlinnetwork.org/`. Accessed: 2025-08-27.

[33] *Beatoven.ai*. `https://www.beatoven.ai/`. Accessed: 2025-08-27.

[34] *Musical AI*. `https://www.musical.ai/`. Accessed: 2025-08-27.

[35] *Splice: Music Creation and Collaboration Platform*. `https://splice.com/`. Accessed: 2025-08-27.

[36] *Moises: AI Music Editing Platform*. `https://moises.ai/`. Accessed: 2025-08-27.

[37] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[38] Y. Mao et al. Multi-RAG: A Multimodal Retrieval-Augmented Generation System for Adaptive Video Understanding. In: *arXiv preprint arXiv:2505.23990* (2025).

[39] Z. Shi et al. VisRAG: Vision-based Retrieval-augmented Generation on Multi-modality Documents. In: *arXiv preprint arXiv:2410.10594* (2024).

[40] C. Zhu et al. Cross-modal RAG: Sub-dimensional Retrieval-Augmented Text-to-Image Generation. In: *arXiv preprint arXiv:2505.21956* (2025).

[41] M. Huang, P. Zhang, and B. Zhang. MotionRAG-Diff: A Retrieval-Augmented Diffusion Framework for Long-Term Music-to-Dance Generation. In: *arXiv preprint arXiv:2506.02661* (2025).

[42] J. S. Downie. The Scientific Evaluation of Music Information Retrieval Systems: Foundations and Current Issues. In: *Journal of the American Society for Information Science and Technology* **54**.12 (2003), pp. 1138–1150.

[43] P. Primus, F. Schmid, and G. Widmer. TACOS: Temporally-aligned Audio CaptiOnS for Language-Audio Pretraining. In: *arXiv preprint arXiv:2505.07609* (2025).

[44] Z. Wang, X. Xia, X. Zhu, and L. Xie. U-sam: An audio language model for unified speech, audio, and music understanding. In: *arXiv preprint arXiv:2505.13880* (2025).

[45] G. Wijngaard, E. Formisano, M. Esposito, and M. Dumontier. AudSemThinker: Enhancing Audio-Language Models through Reasoning over Semantics of Sound. In: *arXiv preprint arXiv:2505.14142* (2025).

[46] S. Alshammari, J. Hershey, A. Feldmann, W. T. Freeman, and M. Hamilton. I-con: A unifying framework for representation learning. In: *arXiv preprint arXiv:2504.16929* (2025).

[47] G. Mariani, I. Tallini, E. Postolache, M. Mancusi, L. Cosmo, and E. Rodolà. Multi-source diffusion models for simultaneous music generation and separation. In: *arXiv preprint arXiv:2302.02257* (2023).

[48] Z. Xu, D. Dutta, Y.-L. Wei, and R. R. Choudhury. Multi-source music generation with latent diffusion. In: *arXiv preprint arXiv:2409.06190* (2024).

[49] T. Karchkhadze, M. R. Izadi, and S. Dubnov. Simultaneous music separation and generation using multi-track latent diffusion models. In: *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2025, pp. 1–5.

[50] T. Karchkhadze, M. R. Izadi, K. Chen, G. Assayag, and S. Dubnov. Multi-track musicldm: Towards versatile music generation with latent diffusion model. In: *International Conference on ArtsIT, Interactivity and Game Creation*. Springer. 2024, pp. 76–91.

[51] D. Yu, K. Song, P. Lu, T. He, X. Tan, W. Ye, S. Zhang, and J. Bian. Musicagent: An ai agent for music understanding and generation with large language models. In: *arXiv preprint arXiv:2310.11954* (2023).

[52] S. Doh, K. Choi, and J. Nam. TalkPlay-Tools: Conversational Music Recommendation with LLM Tool Calling. In: *arXiv preprint arXiv:2510.01698* (2025).

[53] UNESCO. *Generative AI and the Future of Creativity*. `https://unesdoc.unesco.org/ark:/48223/pf0000386088`. Accessed: 2025-08-29. 2023.

[54] MIDiA Research. *The Power of Superfans: How the Most Engaged Fans Drive Music's Future*. `https://www.midiaresearch.com/reports/the-power-of-superfans`. Accessed: 2025-08-29. 2021.

[55] McKinsey & Company. *The Economic Potential of Generative AI*. `https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai`. Accessed: 2025-08-29. 2023.

[56] D. Hesmondhalgh. *Why Music Matters: Resisting the Post-Streaming Economy*. Polity Press, 2023.

[57] L. M. Meier. Post-Streaming Music Cultures: Rethinking Music Economies Beyond Platforms. In: *Popular Music and Society* **46**.4 (2023), pp. 389–407.

[58] Anthropic. *Model Context Protocol (MCP)*. `https://modelcontextprotocol.io/`. Accessed: 2025-08-27. Nov. 2024.

A Music AI Agent Pipeline: Architecture and Workflow

A.1 System Overview

This system is an agent-based platform engineered for interactive music generation and understanding, operating on a human-in-the-loop paradigm. The system's architecture is founded upon the synergistic integration of four core pillars: (1) a structured, modular music database, termed **BlockDB**, which serves as the foundational layer for content retrieval; (2) a versatile suite of **generative and analytical music tools** for processing and creating a wide range of musical information; (3) a sophisticated **session-based multi-agent system** that orchestrates the creative workflow by interpreting user intent and managing task execution; and (4) a real-time, automated **Attribution Layer** that ensures transparent and equitable royalty distribution.

Our architecture prioritizes iterative collaboration, granular control, and transparent attribution. Unlike monolithic models that generate entire tracks from a single prompt, our agent adopts a modular, session-based approach. This method treats music as a structured assembly of components, mirroring how human composers often layer elements like rhythm and melody sequentially. This modularity not only enables a more interactive user experience but also makes complex tasks more controllable by breaking them into smaller, constrained problems, such as generating a single instrumental part. Critically, this session-based paradigm is the foundation for the system's robust attribution mechanism.

A.2 BlockDB: A Granular Music Component Database

A.2.1 The *Block* as a Fundamental Unit

The atomic unit of musical information within the ecosystem is the "Block." Each Block represents a discrete audio segment of a single instrumental or vocal stem, situated within a specific structural section of a composition. To formalize this concept and enable systematic organization and retrieval, Blocks are categorized along a two-dimensional ontological framework:

- **Timbral axis:** This axis classifies Blocks based on their timbral characteristics and instrumental role. For example, a Block could be a distinct instrumental stem such as a percussive drum loop, a bassline played by a bass guitar, a chord progression from a piano, or a melodic synth lead.
- **Temporal axis:** This axis provides temporal context by classifying Blocks according to their typical position within a song structure, such as an intro, a verse, or a chorus.

This dual-axis categorization imposes a formal structure on musical ideas, transforming a potentially disorganized collection of audio files into a structured repository. This structure functions as a compositional grammar, allowing the AI to reason about abstract musical concepts such as "a high-frequency melodic element for a chorus" as a distinct, queryable object. This moves beyond simple timbre matching to a system that understands compositional context, which is a significant step towards more musically aware AI.

A.2.2 Enabling Retrieval-Augmented Generation (RAG) and Attribution

Every Block stored in the database is an atomic entity containing not only the audio data but also a rich set of metadata. This metadata comprises several key components: (1) musical properties, including the Block's timbral and temporal type, BPM, key, and bar length; (2) descriptive text captions; and (3) a suite of embeddings for various modalities. Crucially, the creator of each Block is meticulously recorded. This structured approach facilitates sophisticated **Retrieval-Augmented Generation (RAG)** and enables robust **attribution tracking** throughout the creative process. When a Block is retrieved, its information can be used in diverse ways; for instance, its audio can serve as a direct condition for a generation task, or it can be transcribed into a symbolic format like MIDI to provide a melodic condition. In all cases, the retrieval triggers attribution to the original creator.

A.2.3 Artist Submissions and Content Ingestion

BlockDB is designed as an open ecosystem that grows through artist contributions, which are ingested via two primary pathways. First, artists can submit finished tracks, which our system then processes through a decomposition pipeline. This pipeline analyzes and decomposes the audio both timbrally into its constituent stems (e.g., vocals, drums, bass) and temporally into structural sections (e.g., verse, chorus). Each resulting component is indexed as a Block with a clear link to the original work, supporting direct attribution. Second, artists and producers can contribute standalone stems

and samples directly. This pathway is crucial for incorporating pre-cleared musical components and addressing complex copyright frameworks. All submissions, regardless of type, undergo a systematic ingestion process to be seamlessly integrated as new Blocks, creating a legally robust and creatively diverse library for the platform.

A.3 A Unified Framework for Generative and Analytical Tools

A.3.1 A Diverse and Extensible Toolset

The agent's core capability is its ability to orchestrate a diverse collection of specialized AI models and conventional signal processing or rule-based tools to accomplish complex user requests. Instead of relying on a single monolithic model, the system leverages a toolset where each tool is optimized for a specific music-related task. This modular toolkit can include, but is not limited to:

- **Generation Tools:** A wide array of models for tasks such as `single-song-generation` from multi-modal inputs (text, symbolic, audio), `stem-generation` which creates a new instrumental track from context and prompt audio, `lyric-to-melody`, `singing-voice-synthesis` for creating vocals from lyrics and scores, and `singing-voice-conversion` for altering the timbre of an existing vocal track.
- **Understanding Tools:** Tools designed to interpret and analyze musical content, including models for tasks like `music-source-separation`, `MIDI-transcription`, `music-classification`, `music-captioning`, and `music-structure-analysis`.

This approach allows the agent to be highly extensible and adaptable. New tools can be integrated into the framework to expand its capabilities without requiring changes to the core architecture.

A.3.2 Generalized Conditional Framework

The power of the agent lies in its ability to dynamically chain these tools, using the output of one as the input for another. This process is guided by a generalized, multi-modal conditioning framework. The agent can process and generate various forms of musical information, including:

- **Text:** Descriptive prompts, lyrics, genre tags, or emotional descriptors.
- **Symbolic Music:** MIDI or other sheet music formats representing melody, harmony, and rhythm.
- **Audio:** Raw waveforms containing timbral, stylistic, and performance characteristics.

This framework allows for complex and creative workflows. For example, the agent can perform `stem-generation` by using an existing mix Blocks as context audio and a short instrumental clip Block as prompt audio to add a new, coherent layer to the track. In another common workflow, a user could provide lyrics (text), which the agent uses with a `lyric-to-melody` tool to generate a MIDI file (symbolic). This symbolic data, along with the lyrics, can then be fed into a `singing-voice-synthesis` model to produce a complete vocal track. This multi-modal control over different musical aspects offers significant creative flexibility.

A.4 The Music AI Agent: An Interactive and Sequential Workflow

As depicted in Figure 4, the system employs a sequential, interactive methodology orchestrated by a multi-agent architecture. This architecture acts as the central controller for all user-facing tasks, managing user interaction, resource retrieval, and tool execution within an iterative, human-in-the-loop process.

1. Intent Agent: Devising a Task Plan The workflow begins when a user submits a prompt. Intent Agent is the first to process this input, analyzing the user's high-level creative goal to devise a direct task plan. This plan is a dependency graph that specifies which tools from the toolset to use and in what sequence. The agent determines how data should flow between these tools to achieve the desired outcome. For example, for a complex request like, "create a new retro-pop synth track with my scratch melody audio over my drum track," the Intent Agent constructs a multi-tool plan:

1. First, execute a `transcription` tool to convert the user's scratch melody audio into a symbolic MIDI file.
2. Second, retrieve a suitable synth sound from BlockDB to serve as the prompt audio for the timbre.

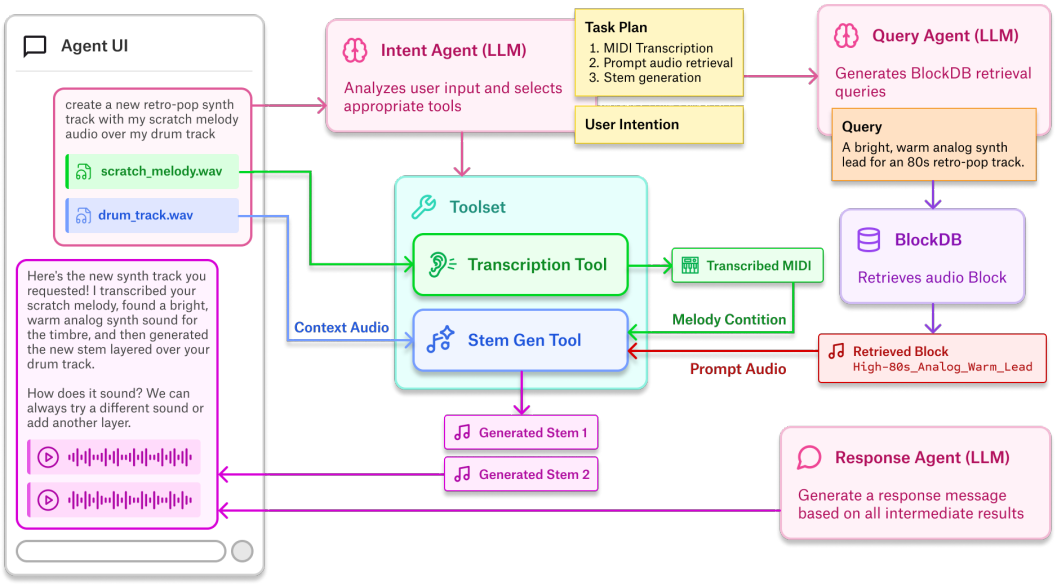

Figure 4: This diagram shows a multi-agent AI's workflow for creating music, where it analyzes a user's prompt and audio files, retrieves a relevant sound Block from BlockDB, and uses various tools to generate new musical stems as a final output.

3. Third, execute a `stem-generation` tool, feeding it three distinct inputs: the transcribed MIDI from the first step as the melodic guide, the user's uploaded drum track as the context audio, and the retrieved synth sound from the second step as the prompt audio.

2. Query Agent: Retrieving Attributable Assets Following the execution plan from Intent Agent, Query Agent is responsible for gathering the necessary creative assets from **BlockDB**. This is the retrieval step in the system's Retrieval-Augmented Generation (RAG) process, where the agent translates abstract requirements from the plan into concrete queries. Continuing the example, to fulfill the plan's need for a "retro-pop synth sound," the Query Agent formulates a query to find a suitable audio Block. The retrieved Block is then designated to serve as the prompt audio for the `stem-generation` tool. This retrieval is the critical link to attribution; the act of fetching a Block immediately triggers a **Block usage event** to be sent to the Attribution Layer, ensuring the original creator is logged for compensation.

3. Task Execution and Session-based Iterative Loop With the execution plan set and all necessary conditional information gathered (from the user or retrieved by the Query Agent), the system executes the specified tools in sequence. The output of one tool, such as the MIDI from the `transcription` tool, seamlessly becomes an input for the next. The final output, such as newly generated stems, is packaged with a natural language **response** explaining the process and sent back to the user. This completes one cycle of the interactive loop, allowing the user to provide further prompts to refine the output, add new musical layers, or initiate a different task, creating a collaborative, session-based, human-in-the-loop workflow.

A.5 Attribution Layer

A core principle of this ecosystem is that attribution is not an afterthought but an integral part of the interactive creation process. The system is designed to track and assign credit in real-time. During the interactive loop, every time a Block is retrieved from BlockDB to serve as a condition for a task—whether as direct audio, a source for transcription, or any other form of creative inspiration—the system immediately logs an attribution event for the original creator of that Block. This ensures that artists are credited for the influence and utility of their contributions at the moment of use, creating a direct incentive for populating BlockDB with high-quality, foundational musical ideas (Figure 5).

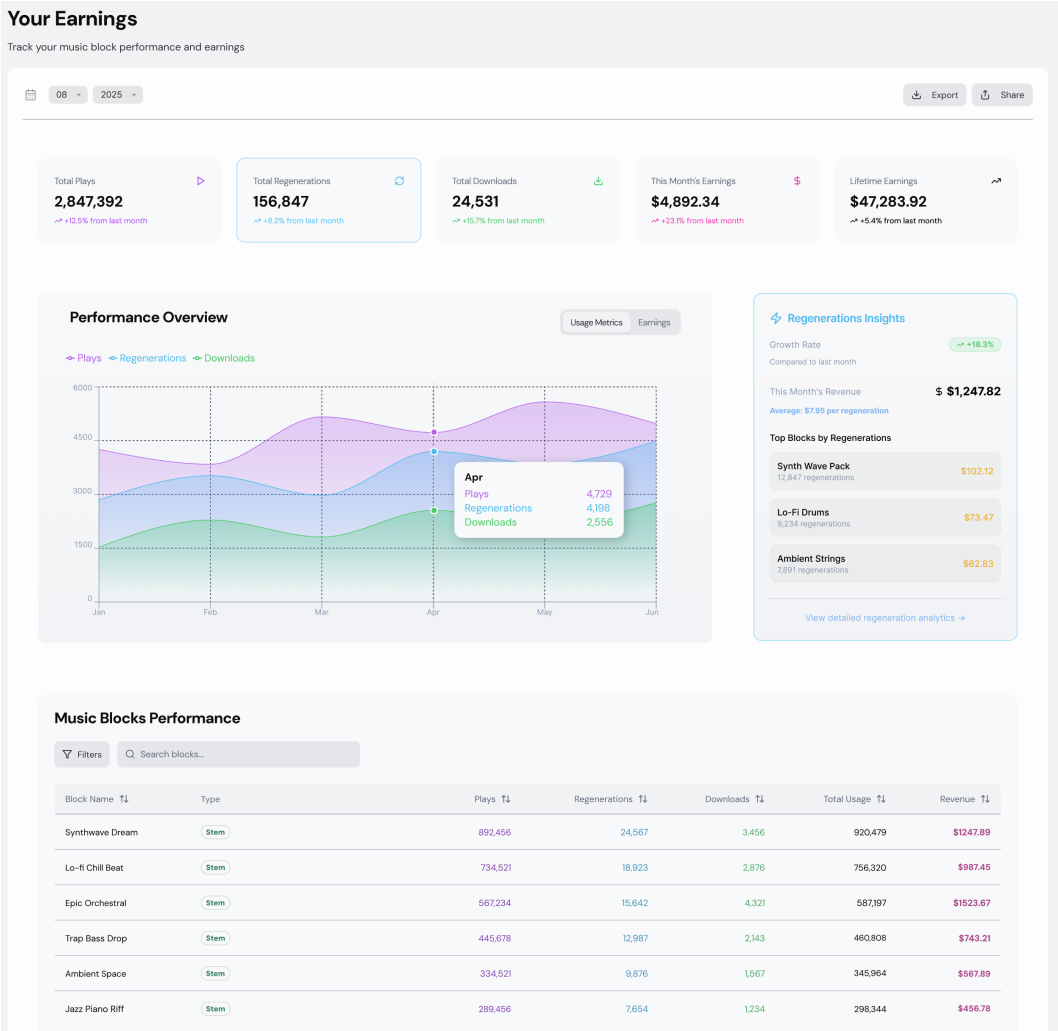

Figure 5: The user interface of dashboard for artists. The dashboard provides detailed analytics on how their individual music Blocks are being used and the revenue they are generating.

A.6 System Extensibility via Model Context Protocol (MCP)

The system is designed for interoperability with the broader digital music ecosystem. Through Model Context Protocol (MCP) [58], the platform can interface with external environments such as Digital Audio Workstations (DAWs), enabling a seamless and extended creative workflow.