

Unified Semantic Representation of Music Scores, Lyrics, and Singing Voices for Multimodal Artificial Intelligence

Gohan Gandhi

Department of Computer Science, University of New Hampshire, Durham, NH, USA.
gandhi23@unh.edu

Arjun M. Srivastava

Department of Computer Science, University of North Texas, Denton, TX, USA.
srivastavaarjun@unt.edu

Abstract

The computational study of music has long been fragmented across symbolic scores, textual lyrics, and audio recordings of singing voices. This fragmentation limits the capacity of artificial intelligence systems to reason across musical modalities, to align lyrical semantics with melodic structure, and to generate expressive performances from combined symbolic and linguistic intent. This paper examines the problem of constructing unified semantic representations for music scores, lyrics, and singing voices within multimodal artificial intelligence. Rather than proposing a single model, the paper develops a system-level account of the structural trade-offs involved in representing discrete symbolic events, continuous acoustic signals, and natural language under a shared semantic framework. The analysis addresses architectural strategies based on attention mechanisms, self-supervised audio representations, and sequence-to-sequence modeling, and it situates these strategies within data infrastructure, annotation governance, and deployment contexts. Particular attention is given to alignment granularity, temporal hierarchy, modality-specific fidelity, robustness across genres and languages, fairness, sustainability, and policy implications. The paper argues that meaningful progress requires not only technical integration but also careful governance of training corpora, evaluation protocols, and downstream sociotechnical effects. A unified representation should be understood as an evolving infrastructure rather than a static embedding space.

Keywords

unified semantic representation; multimodal artificial intelligence; music information retrieval; singing voice synthesis; score-lyrics alignment; data governance.

1. Introduction

The development of computational systems for music has produced capable but largely disconnected tools for symbolic score analysis, lyric processing, and singing voice synthesis. These modalities differ in their underlying representational assumptions. A music score encodes discrete events such as notes, rests, meter changes, and dynamic markings. Lyrics encode phonemic, semantic, and prosodic information through natural language. A singing voice recording encodes a continuous acoustic signal shaped by phonation, articulation, vibrato, and expressive timing. Because each modality has been addressed within its own research community, current artificial intelligence systems often treat score, text, and audio as

separate pipelines that are later combined through brittle interfaces. Representation learning has demonstrated the value of distributed representations that capture underlying regularities in complex data [1]. Attention-based sequence models have reshaped machine translation, speech processing, and symbolic music generation by allowing flexible alignment between sequences [2]. Transfer learning has enabled representations trained on large text corpora to be adapted to many downstream language tasks [3], while self-supervised speech models have produced increasingly robust acoustic representations without extensive manual labels [4]. In symbolic music generation, architectures designed for long-range structure have shown that symbolic event sequences can be modeled with considerable coherence [5]. These advances create an opportunity to ask whether a unified semantic representation for music scores, lyrics, and singing voices can serve as a foundation for multimodal musical intelligence.

The motivation for such a representation is not merely technical convenience. A composer, performer, and listener do not experience score, text, and voice as independent channels; rather, they integrate them into an interpretable musical meaning. A unified representation would need to preserve the specific logic of each modality while enabling cross-modal alignment, retrieval, transformation, and generation. This paper examines that challenge from a system-level perspective. It considers the structural trade-offs between discrete and continuous representations, the architectural choices for aligning heterogeneous sequences, the design of multimodal data infrastructures, and the governance mechanisms required for responsible deployment. The argument advanced here is that unified semantic representation should be treated as a sociotechnical infrastructure problem rather than a narrowly defined model architecture problem. Such a perspective foregrounds questions of annotation quality, fairness across musical traditions, robustness under noisy audio conditions, sustainability of large-scale training, and policy constraints on data use.

2. Conceptual Foundations and Motivations for Unified Semantic Representation

Music scores, lyrics, and singing voices encode musical meaning in fundamentally different ways. A score is a symbolic abstraction that specifies pitch, duration, onset, and intensity, but it does not fully determine how a performance should sound. Lyrics are language objects with phonological, syntactic, semantic, and poetic dimensions, while a singing voice is an acoustic realization that combines phonetic content with melodic and expressive variation. The semantic units in each modality are not naturally aligned: a syllable may span several notes, a note may be ornamented without explicit notation, and a sung vowel may carry emotional meaning that is absent from both the score and the text. A unified semantic representation must therefore capture correspondences at multiple temporal granularities while allowing modality-specific information to remain available when needed.

Recent efforts in multimodal music data collection have begun to provide paired scores, lyrics, and singing recordings at larger scales, enabling systematic study of cross-modal consistency [6]. These resources illustrate a broader shift from isolated tasks, such as score following or lyric transcription, toward integrated tasks that demand simultaneous reasoning about symbolic, linguistic, and acoustic structure. At the same time, learned audio representations in music information retrieval have shown that convolutional and recurrent models can capture timbre, genre, and instrument identity, although such representations often remain at the segment level rather than explicitly aligned with score events or phonemes [7]. The challenge is not simply to concatenate these representations but to define semantic anchors that are invariant enough for cross-modal transfer and specific enough for expressive control.

A unified representation would have several practical motivations. It could support score-informed singing voice synthesis, where the model must respect both notated pitch and lyrical intelligibility. It could improve lyric-to-melody generation by grounding language in musically plausible pitch contours. It could enable query-by-humming systems that retrieve scores or lyrics from sung fragments. It could also assist music education by providing aligned visual, textual, and auditory feedback. However, these applications depend on a deeper conceptual clarification of what should be shared across modalities. A representation that discards all acoustic detail may be useful for symbolic retrieval but inadequate for expressive synthesis. Conversely, a representation that preserves every acoustic nuance may be too entangled with performer identity to support symbolic manipulation. The central theoretical problem is therefore to identify which aspects of musical meaning are genuinely cross-modal and which aspects should remain modality-specific.

3. Multimodal Representation Challenges and Structural Trade-offs

The construction of unified semantic representations for score, lyrics, and voice involves several interacting structural trade-offs. One major challenge concerns the relationship between discrete symbolic tokens and continuous acoustic embeddings. Self-supervised speech representation learning has shown that acoustic segments can be mapped to discrete units through clustering, and that these units often carry strong phonetic content even without transcriptions [8]. This finding suggests that continuous audio can be transformed into symbolic-like representations through data-driven learning. However, music audio includes additional dimensions such as pitch, harmony, timing, and timbre that do not always align cleanly with phoneme-like units. Generative models of music audio have demonstrated that raw waveforms can be modeled using hierarchical discrete codes with conditioning on lyrics and musical attributes [9]. These models can produce convincing outputs, but they do not necessarily learn explicit score-to-voice alignment, and their internal representations may remain opaque.

A second structural trade-off concerns temporal granularity. A score event sequence is relatively sparse, with notes lasting hundreds of milliseconds or more. Lyric phonemes are denser, and audio frames are denser still. Any shared representation must reconcile these different temporal resolutions. A common strategy is to define multiple levels of abstraction, such that high-level units correspond to phrases or measures, intermediate units correspond to notes or syllables, and low-level units correspond to acoustic frames. This hierarchical organization can reduce alignment ambiguity, but it also introduces additional design decisions about how information should propagate across levels. If the hierarchy is too rigid, it may fail to capture melismas, ornaments, and expressive timing. If it is too flexible, the representation may lose the symbolic interpretability that makes it useful for music analysis.

A third trade-off involves the geometry of the shared semantic space. When score, lyrics, and voice are mapped into a common latent space, there is a risk of collapsing information that should remain separable. For instance, singer identity, pronunciation accent, and emotional expression are important in some tasks but irrelevant in others. A representation optimized for lyric-to-score alignment may suppress these properties, while a representation optimized for high-fidelity synthesis may preserve them at the expense of cross-modal invariance. The appropriate balance depends on the intended downstream use. System designers must therefore decide whether to use a single shared encoder, separate unimodal encoders with shared alignment layers, or a hybrid in which shared and modality-specific subspaces coexist. Each choice has implications for controllability, interpretability, and data efficiency.

4. Architectural Strategies for Score-Lyrics-Voice Integration

Recent architectural developments provide several building blocks for multimodal music representation. Symbolic scores can be tokenized as event-like sequences that include note onset, duration, pitch, and score position, enabling sequence models to process them in a manner similar to language [10]. This tokenization must be coordinated with lyric tokenization and audio tokenization. Music processing research has long established principles for representing notes, beats, and pitch in ways that support alignment and retrieval [11]. The choice of token vocabulary is not neutral; it defines which musical features are treated as first-class objects and which are treated as residual noise. For example, a vocabulary that lacks explicit ornaments cannot easily represent melismatic singing even if the acoustic model is powerful.

Sequence-to-sequence architectures offer a natural framework for integrating multiple modalities. Large-scale audio-text models have shown that audio and language can be aligned through encoder-decoder training with weakly labeled data, producing strong transfer to speech recognition and translation tasks [12]. Speech synthesis architectures have similarly demonstrated that text can be mapped to acoustic features using explicit duration predictors and pitch modules [13]. For unified music representation, a comparable architecture could condition on both score and lyrics to generate a singing voice. The decoder would need to respect notated pitch and rhythm while also producing intelligible phoneme sequences. Cross-attention mechanisms are central to this process because they allow the decoder to query relevant score positions and lyric phonemes at different points in time. However, the alignment problem is more complex than in ordinary speech synthesis because the score provides an additional symbolic constraint that may conflict with natural lyric stress or expressive phrasing.

Training objectives also shape the semantic properties of the resulting representation. Reconstruction objectives encourage the model to preserve acoustic detail, while contrastive objectives encourage cross-modal invariance. Adversarial objectives can improve perceptual quality but may introduce instability. Multitask learning can combine several objectives, but it can also produce optimization conflicts if the tasks impose incompatible preferences on shared parameters. A promising approach is to use progressive training, in which a model first learns unimodal representations and then learns cross-modal alignment through a shared semantic bottleneck. This kind of staged training can reduce the difficulty of learning from heterogeneous data and can provide more interpretable intermediate representations. Nonetheless, it requires careful design to avoid catastrophic forgetting and to ensure that the shared representation retains musical meaning across all three modalities.

5. Data Infrastructure, Annotation, and Semantic Governance

Unified multimodal music representation depends heavily on the availability of high-quality aligned data. Annotation schemas must cover musical notation, lyric syllabification, phoneme alignment, note boundaries, expressive markings, and possibly singer metadata. Recent corpus efforts with paired scores and realistic singing performances provide a useful reference for alignment research [6]. However, any large dataset is the result of many human decisions about what to annotate, how to resolve ambiguities, and which musical traditions to include. These decisions can shape model behavior in ways that are not always visible during evaluation. Model cards offer a structured way to document intended use, limitations, training data, and performance characteristics [14]. Datasheets extend this logic to datasets by recording collection procedures, annotation protocols, demographic distributions, and

licensing information [15]. Both instruments are highly relevant to multimodal music data because they can make hidden assumptions legible to downstream users.

The governance of musical data also raises distinctive legal and ethical issues. Music involves multiple rightsholders, including composers, lyricists, performers, and producers. Licensing regimes vary across jurisdictions, and the rights associated with a score may differ from those associated with a particular recording. Singing voice data are especially sensitive because the voice is a biometric identifier and a carrier of personal identity, health state, and cultural expression. A unified representation trained on singing voice recordings may inadvertently encode speaker-specific features even when the task requires only musical or linguistic alignment. This creates privacy risks and can also reduce generalization if the system overfits to repeated singer identities. Annotation governance should therefore include singer consent for model training, clear statements about intended reuse, and technical measures such as disentangled representation learning that prevent unauthorized voice reconstruction. The broader point is that the semantic value of a multimodal music representation cannot be separated from the social relations through which its training data were created.

6. Robustness and Fairness in Cross-Modal Inference

A unified semantic representation for scores, lyrics, and singing voices must remain robust under acoustic variation, annotation noise, and stylistic diversity. Studio-quality singing recordings with matched scores provide a controlled starting point, but deployed systems frequently encounter amateur singing, background noise, reverberation, dialectal pronunciation, and nonstandard notation. Representations trained predominantly on clean Western concert singing may exhibit brittle alignment when faced with speech-like singing, extreme vibrato, or pitch ornamentation from oral traditions. Music information retrieval research has long noted that many widely used datasets implicitly privilege Western popular and classical music, which limits the cultural generality of learned models [16]. Addressing this limitation requires deliberate corpus expansion across languages, singing styles, and recording conditions rather than assuming that scale alone will resolve representational bias. The tendency of large-scale language and speech models to overrepresent dominant languages and speaker populations has been widely documented, and similar dynamics are likely to emerge in unified music models unless explicitly counteracted [17].

Fairness concerns in multimodal music systems arise in several forms. First, the phonemic and prosodic structure of lyrics varies across languages, and many singing voice models inherit phoneme inventories optimized for English or other high-resource languages. This can produce unintelligible output for tonal languages or for languages with rare phonemes. Second, the symbolic score conventions used in training are often tied to Western staff notation, which may not capture microtonal pitch systems or rhythmic cycles in non-Western musics. Third, singer identity and accent may be entangled with musical meaning in ways that a representation either preserves or erases. A representation that suppresses singer identity may improve privacy but may also remove culturally meaningful aspects of vocal timbre. A representation that retains biometric information can expose individuals to unauthorized voice synthesis. These trade-offs require explicit design decisions and transparent documentation. The literature on model cards and datasheets offers useful mechanisms for recording such choices and their consequences for protected groups and minority musical traditions [14], [15].

Robustness also depends on the alignment inference process. When a system aligns a lyric syllable to a notated note and a sung frame, errors can arise from ambiguous segmentation, missing score annotations, or cross-modal timing drift. A representation that relies on monotonic alignment assumptions may fail for melismatic ornamentation that does not follow note-syllable correspondences. In contrast, a representation that permits arbitrary long-range dependencies may be more flexible but also more difficult to train and less interpretable. This is a classic structural tension between expressiveness and controllability. Mitigating it requires not only architectural innovation but also evaluation protocols that measure performance separately across annotation types, musical genres, and noise conditions.

7. Deployment, Sustainability, and Computational Infrastructure

The construction and deployment of unified semantic representations raise substantial computational sustainability questions. Large-scale multimodal pretraining often demands thousands of GPU-hours, and singing voice audio processing can be particularly costly because of the need for high temporal resolution and large batch sizes. The environmental and financial costs of such training have prompted broader debates about energy use in machine learning research [18]. Unified music representations, if designed without efficiency constraints, may repeat the pattern of increasingly large models whose benefits accrue primarily to well-resourced institutions. This is especially problematic for music, which is a globally distributed cultural practice that should not be reduced to what can be affordably modeled in high-income academic labs.

There are several structural responses to this challenge. One is to develop modular architectures that allow unimodal components to be pretrained independently before lightweight cross-modal alignment. This reduces the need to retrain large audio and text encoders for every new alignment task. Another response is to use shared latent spaces with adjustable capacity depending on deployment context, so that a research system can use high-capacity representations while edge devices use quantized or distilled variants. The concept of green AI emphasizes efficiency as a first-class research objective rather than an afterthought, and this principle applies directly to multimodal music systems [19]. A third response is to invest in data-efficient representation learning that exploits paired score, lyric, and audio modalities as a form of mutual supervision, thereby reducing dependence on large volumes of manually annotated data. However, efficiency must not come at the cost of fairness or robustness, because a smaller model may overfit to dominant musical styles and fail to generalize to minority traditions.

Deployment also involves real-time constraints. Singing voice processing in interactive applications, such as live score following or voice transformation, requires low-latency alignment and incremental inference. Unified representations may need to function in both batch and streaming modes, which places additional pressure on the temporal hierarchy of the shared space. Systems designed for offline music analysis may assume access to a full score and full audio file, but deployed applications often receive partial inputs. This mismatch between training-time and inference-time assumptions is a significant systems engineering challenge that should be addressed through careful interface specification and through benchmarks that simulate streaming conditions.

8. Evaluation, Benchmarks, and Reproducibility

The evaluation of unified semantic representations for score, lyrics, and voice is complicated by the absence of a single task-neutral metric. A representation may excel at lyric-to-score

alignment but perform poorly at expressive singing synthesis or at preserving linguistic intelligibility. Current benchmarks often focus on one modality or one interface, such as note transcription accuracy or mean opinion scores for voice naturalness. Symbolic music pretraining has produced benchmarks for masked token prediction and classification, but these do not necessarily test cross-modal semantic consistency [20]. Large multimodal speech and audio models have been evaluated through task-specific leaderboards, yet similar infrastructures for music remain fragmented [12]. A more mature evaluation ecosystem would include paired tasks that require simultaneous reasoning across all three modalities, such as retrieving a score from a sung phrase, generating a singing voice from an incomplete score annotation, or detecting mismatches between lyric stress and notated meter.

Reproducibility is another persistent concern. Many music and speech models are released without full training code, and the preprocessing of score, lyric, and audio data can introduce hidden variations that affect alignment accuracy. The machine learning community has developed reproducibility frameworks that emphasize the importance of documenting data selection, hyperparameter sensitivity, and evaluation methodology [21]. For multimodal music representation, reproducibility also requires the preservation of annotation provenance. If a score encoding tool changes its tokenization scheme, or if a phoneme alignment package is updated, downstream representations may shift in ways that invalidate prior comparisons. Versioned data artifacts and stable evaluation harnesses are therefore essential components of the research infrastructure.

Human evaluation raises additional methodological issues. Musical quality is subjective and culturally conditioned, and listeners may disagree about the acceptability of synthesized singing voices. Standardized listener panels can provide useful signals but may also reflect particular aesthetic preferences unless panel design is carefully diversified. Furthermore, human judgments of lyric intelligibility and emotional expressiveness can be influenced by a singer’s accent, which may introduce bias against minoritized speech communities. Addressing these concerns requires transparent reporting of evaluator demographics and the use of multiple complementary evaluation methods rather than a single aggregate score.

9. Policy Implications and Governance

The development of unified representations that can generate or manipulate singing voices has significant policy implications. Voice cloning and deepfake audio have raised widespread concern about deception, consent, and identity misuse [22]. Singing voices are especially vulnerable because musical performances often carry strong emotional and cultural associations. A model that can synthesize a recognizable singer’s voice from a small corpus could be used to create unauthorized recordings, which may violate publicity rights and copyright law in many jurisdictions. Regulatory frameworks for artificial intelligence, such as emerging obligations for transparency and risk management, increasingly address biometric data and generative audio systems [23]. However, the specific intersection of musical works, performer rights, and voice biometrics remains legally unsettled in many parts of the world.

Governance should not be limited to legal compliance. Researchers and system deployers have an ethical responsibility to consider how unified multimodal music representations may affect working musicians, vocalists, and communities whose musical traditions are used as training data. A representation that enables high-fidelity style transfer may disrupt livelihood by reducing demand for human vocal performance in certain production pipelines. Conversely, the same technology may expand access by enabling composers who do not sing to produce convincing vocal demos. These potential benefits and harms are not uniformly distributed,

and policy mechanisms should therefore be developed through participatory processes that include musicians, rights organizations, linguistic communities, and technology developers. Documentation artifacts such as datasheets and model cards can make some of these considerations transparent, but they do not replace substantive consultation [15], [14].

Data governance itself is an infrastructure problem. Multimodal music datasets often combine scores, lyrics, and audio from different sources with incompatible licenses and varying levels of documentation. A unified representation may depend on a delicate integration of these sources, and any downstream model inherits the legal status of its training data. FAIR guiding principles for scientific data management provide a useful starting point for making such datasets findable, accessible, interoperable, and reusable, though they must be supplemented by rights-aware practices specific to music [24]. The future of unified semantic representation depends as much on sustainable and lawful data pipelines as on machine learning architecture.

10. Conclusion

Unified semantic representation of music scores, lyrics, and singing voices represents a promising but underexplored frontier for multimodal artificial intelligence. This paper has argued that the central challenge is not merely to combine three data modalities into a single neural network, but to develop a system-level framework that respects the distinct symbolic, linguistic, and acoustic logics of each modality while enabling meaningful cross-modal inference. The structural trade-offs are substantial. Discrete score events and continuous singing audio require careful temporal and geometric alignment. Representation spaces must balance modal invariance against expressive specificity. Training pipelines must reconcile the constraints of limited aligned data, diverse musical traditions, and computational sustainability. Governance structures must address privacy, consent, intellectual property, and cultural representation.

A unified semantic representation is best understood as an evolving sociotechnical infrastructure. It includes tokenization schemes, alignment protocols, training objectives, evaluation benchmarks, documentation standards, and policy mechanisms. The technical architecture cannot be separated from the annotation decisions, licensing practices, and fairness considerations that shape the data. Future research should move beyond single-model showcases toward longitudinal system studies that examine how such representations behave when deployed across different musical contexts, languages, and recording conditions. This requires interdisciplinary collaboration among machine learning researchers, music theorists, linguists, sound engineers, legal scholars, and musicians themselves. The goal is not a universal embedding that erases difference, but a shared semantic framework that supports informed translation, creative transformation, and responsible interaction across score, lyric, and voice.

References

1. Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798–1828.
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.
4. Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33, 12449–12460.
5. Huang, C. Z. A., Vaswani, A., Uszkoreit, J., Simon, I., Hawthorne, C., Shazeer, N., ... & Eck, D. (2019). Music Transformer: Generating music with long-term structure. *International Conference on Learning Representations*.
6. Zhang, Y., Pan, C., Guo, W., Li, R., Zhu, Z., Wang, J., ... & Zhao, Z. (2024). Gtsinger: A global multi-technique singing corpus with realistic music scores for all singing tasks. *Advances in Neural Information Processing Systems*, 37, 1117-1140.
7. Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). Convolutional recurrent neural networks for music classification. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2392–2396.
8. van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30.
9. Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzett, M., Caillon, A., ... & Frank, C. (2023). MusicLM: Generating music from text. *arXiv preprint arXiv:2301.11325*.
10. Oore, S., Simon, I., Dieleman, S., Eck, D., & Simonyan, K. (2020). This time with feeling: Learning expressive musical performance. *Neural Computing and Applications*, 32, 955–967.
11. Müller, M. (2015). *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer.
12. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning*, 28492–28518.
13. Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. Y. (2021). FastSpeech 2: Fast and high-quality end-to-end text-to-speech. *International Conference on Learning Representations*.
14. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.
15. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92.
16. Schedl, M., Gómez, E., & Urbano, J. (2014). Music information retrieval: Recent developments and applications. *Foundations and Trends in Information Retrieval*, 8(2-3), 127–261.

17. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.
18. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650.
19. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM*, 63(12), 54–63.
20. Zeng, M., Tan, X., Wang, R., Ju, D., Qin, T., & Liu, T. Y. (2021). MusicBERT: Symbolic music understanding with large-scale pre-training. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 791–800.
21. Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., ... & Bengio, Y. (2021). Improving reproducibility in machine learning research: A report from the NeurIPS 2019 reproducibility program. *Journal of Machine Learning Research*, 22(1), 7459–7478.
22. Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107, 1753–1820.
23. Veale, M., & Zuiderveen Borgesius, F. (2022). Demystifying the draft EU artificial intelligence act—analysing the good, the bad, and the unclear elements of the proposed approach. *Digital Society*, 1, 1–24.
24. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3, Article 160018.