Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30683

Evaluating the Robustness of REMI-Tokenized and Regression-Based Symbolic Transformer under MIDI Distortions

Muge Liu

Newport High School, Bellevue, USA
s-liume@bsd405.org

Abstract. The different touch methods on the piano can be identified as performance errors from distinct sound features, but it is still difficult to evaluate them solely through human intervention. Machine learning models can effectively solve this problem. This paper evaluates symbolic transformer model's ability to automatically assess piano performance by utilizing controlled MIDI distortions. Using 1276 MAESTRO v3 excerpts, this research generates 27868 MIDI files categorized into eight distortion classes at three severity levels. This model is a REMI-tokenized regression-based transformer adaptation that is trained for 60 epochs on an 80/20 training and validation split. The regression model attains 63% validation accuracy with a validation loss of 0.28. The model demonstrates various levels of sensitivity to timing and pitch distortions. Results indicate that REMI tokenization and RegressLM integration improve detection of severe and moderate errors but struggle with subtle expressive deviations, highlighting directions for augmentation and curriculum learning to strengthen applicability.

Keywords: Regression-Based Transformer, REMI Tokenization, Music Representation Learning, Symbolic Music Modeling, MIDI Distortion

1. Introduction

Accurate and applicable assessment of piano performance is an essential problem between music education and sequence modeling. Automated evaluation systems provide pedagogical guidance, objective feedback for practice, and new tools for musical analysis. However, creating systems that reliably distinguish true performance errors from expressive interpretation is challenging [1].

Symbolic MIDI effectively simulates musical performance, reduces the acoustic variability that complicates audio-based evaluation, and enables nuanced comparisons, proving them as an ideal approach for automated assessment achieved with modern symbolic models [2].

Recent advances in transformer-based symbolic modeling have improved the capacity to learn long-range musical structure and styles. Multi-task and muti-track formulations such as MT3 and more recent generative transformers (e.g., MuPT, YourMT3+) exploit large symbolic corpora to produce robust transcription and generation performance across instruments and styles [3].

DOI: 10.54254/2755-2721/2026.TJ30683

Complementary tokenization strategies and beat-aware encodings (e.g., REMI), have further improved how symbolic music is represented for transformer training [4]. At the same time, transformer approaches tailored to error detection have shown promise for explicit pedagogical tasks, such as labeling incorrect notes or computing alignment scores that match human judgements [5].

Despite methodological progress, a conspicuous empirical gap remains. Large symbolic models have not yet been exposed to controlled and pedagogically motivated perturbations of MIDI performances. Prior work has produced valuable datasets and perceptual annotation schemes [6-8], and separate studies have developed systems for musical evaluation and error detection [5,9]. Yet there is limited published evidence implementing a standardized distortion regime that mimics realistic student errors, such as timing perturbations and omitted notes, across graded severities. Additionally, this study aims to classify errors without relying on correct audio for reference, further developing autonomous error detection.

This study constructs a reproducible distortion framework applied to 1276 MAESTRO v3 piano samples and evaluates the application of REMI tokenization and a RegressLM adaptation to the scoring task. A standard transformer model is trained using the same dataset, serving as a side-by-side comparison. Building on Google DeepMind's RegressLM implementation as a regression head and training scaffold [10], this adaptation is trained for 60 epochs on an 80/20 training and validation split.

This work consists of four main contributions. Firstly, a controlled distortion protocol for symbolic piano data is released. It contains eight pedagogical error modes at three severity levels. Secondly, an evaluation of transformer model adapted via RegressLM and REMI tokenization is provided, reporting accuracy, loss, and other evaluation metrics. Thirdly, the model's sensitivity to musical dimensions and difficulty to separate between distinct types of error is analyzed. Lastly, this paper discusses practical implications for automated practice tools and outline pathways to improve reliability in pedagogical settings [7,9].

2. Related work

2.1. Symbolic models

Transformer architectures for symbolic music have rapidly evolved from beat-aware generators to large sequence models that support multi-task transcription and generation. Tokenization and beat-aware encodings remain crucial. Beat-aligned representations introduced in the Pop Music Transformer lineage influenced REMI-style encodings that improve positional interpretations for musical events [4].

2.2. Datasets

A large, curated symbolic corpus is suited for modern transformer models. MAESTRO supplies aligned audio-MIDI pairs and expressive pianistic detail [2], while GiantMIDI-Piano and MidiCaps broaden coverage across repertoire and annotation modalities [6,7]. Perceptually annotated evaluation resources, such as the multi-level perceptual dataset [11], provide supervised labels useful for pedagogical validation and model calibration. Due to the flexibility and easy adaptability of MAESTRO with its midi and audio capabilities, this study employs the dataset and manipulates it to easily create distortion classes and analyze artificially created errors with human verification.

DOI: 10.54254/2755-2721/2026.TJ30683

2.3. Piano performance assessment

Prior work on automated assessment ranges from discriminative neural models that evaluate musical shapes to transformer-based scoring and explicit error detection frameworks. Siamese residual networks have been proposed to evaluate phrase-level shape for pedagogy [9], while transformer detectors provide fine temporal alignment and error labeling capabilities [5,12]. Methodologically, text-to-text regression, specifically RegressLM, offers an adaptable pathway to frame scoring as a continuous prediction task [10,13]. The approach of RegressLM is an ideal integration into this model because knowledge of a correct midi file version is not required, which provides flexibility to analyze and classify computer-generated audio without limiting the data to specific midi versions.

3. Methodology

1276 piano excerpts are sampled from MAESTRO v3 and are manipulated and distorted to form 27868 total files in eight classes as follows: clean (no error), tempo distortion (inconsistent BPM), dropped notes (missing notes), extra notes, onset jitter (timing deviation for specific notes), pitch shift (up or down from original pitch), pitch cluster (multiple notes instead of a single note), and pitch substitution (change to a different note) [2]. These distortion classes simulate real student errors. Each class consists of three corruption severity levels (low, medium, high) to ensure sufficient data coverage. MIDI files are converted to REMI tokens for model input, and the set is partitioned into an 80% training and 20% validation split.

The baseline model employs a standard transformer approach without adaptations. The principal adaptation uses Google DeepMind's RegressLM scaffold to produce continuous scoring outputs alongside REMI encoding to allow for cleaner model input [10,13]. The models are trained for 60 epochs each with a batch size of 8, while the adaptation utilizes REMI encoding and AdamW optimizer with a learning rate of 3e-4. The principal adaptation consists of 384 dimensions, 3 layers, and 6 attention heads. It is trained on a device with CUDA to support a large dataset. A comparative graph of training and validation accuracy is reported along with validation loss. Per-distortion sensitivity is shown and analyzed with a confusion matrix.

4. Results

The REMI-tokenized RegressLM adaptation achieved 63% validation accuracy and a validation loss of 0.28. Comparatively, the standard transformer model achieved only 30% accuracy. The adaptation is a significant improvement relative to initial baselines without the implementation of regression and REMI encoder.

DOI: 10.54254/2755-2721/2026.TJ30683

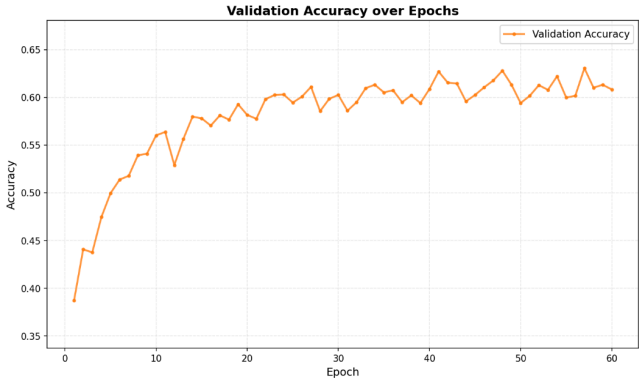

Figure 1. Accuracy of validation dataset over 60 epochs. Validation accuracy shows increase but plateaus at 63%

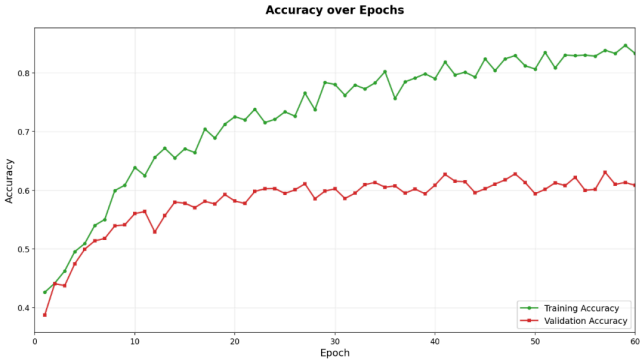

Figure 2. Comparative graph with training and validation accuracy over 60 epochs

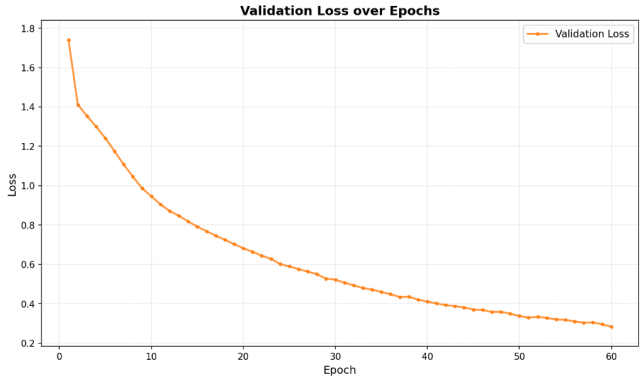

Figure 3. Validation loss over 60 epochs (model shows consistent improvement with steadily increasing certainty in predictions shown by the decrease in loss)

From figure 1-3, although the model's validation accuracy shows an increasing trend and demonstrates clear improvement from the baseline transformer model, its comparison with the training growth indicates model overfitting. Possible causes include overcomplexity of the model and excessive epochs, causing the model to memorize specific properties of specific audio instead of producing generalized predictions. Hyperparameters can be modified to achieve more optimal results. Validation loss consistently decreases, showing steady improvement with the model's increasing certainty in predictions.

DOI: 10.54254/2755-2721/2026.TJ30683

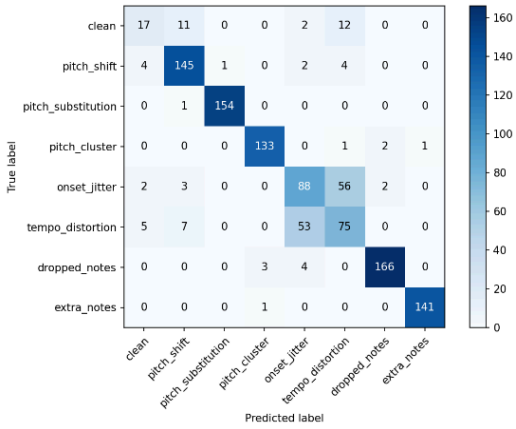

Figure 4. Confusion matrix visualization of epoch 57 training set

Ideally, the confusion matrix forms a colored diagonal line from top left to bottom left, indicating all true positives (figure 4). This set of training data shows around 85% accuracy, which suggests overfitting when compared with 63% accuracy from the validation set at epoch 57. Due to overfitting, the model memorizes specific patterns during training instead of continuously learning more general patterns. However, distinction between onset jitter and tempo distortion remains challenging, demonstrating extremely nuanced and indistinguishable differences between the two distortion classes.

Further general analysis shows that timing perturbations (tempo distortion and onset jitter) are difficult to differentiate, while pitch-level distortions (dropped notes and extra notes) are significantly stronger in accuracy. These patterns are consistent with prior error-detection studies [5,12]. Clean distortion class is rarely predicted correctly, indicating disproportional class weights. Since clean audio consists of all original MIDI samples from MAESTRO v3, its size is one third of other distortion classes. Although severity classes improve data coverage and increase accuracy in these classes, they undermine and outweigh the class with original audio, leading to the model's preference of classifying original audio as distorted classes.

5. Discussion

Results suggest that a custom transformer model with REMI encoding and RegressLM integration yields measurable gains for scoring, yet symbolic models retain blind spots, particularly under severe timing perturbations where similar distortion types are confounded. The low accuracy on clean audio class requires class weight adjustments and more intensive data augmentation to augment low sample size.

Curriculum augmentation that prioritizes gradually increasing distortion severity, along with adversarial augmentation tuned to timing errors, may improve robustness. A 63% automated accuracy indicates utility for high-level feedback (e.g. detecting severe errors) but insufficient reliability for nuanced grading without human oversight. However, classification output remains pedagogically informative. Limitations of this study include a single training and validation split, a finite sample of original MIDI excerpts, and limited types of artificial distortions that may not encapsulate the variability of real student errors.

To model nuance and detect errors more reliably in the future, this study suggests cross-dataset fine-tuning, curriculum learning, and exploring hybrid audio-symbolic approaches. Severity level

DOI: 10.54254/2755-2721/2026.TJ30683

distortions can be developed to serve more than the primary purpose of increasing data coverage and variability.

6. Conclusion

This paper introduces a reproducible systematic distortion framework for symbolic piano data that consists of eight error classes across three severity levels. Additionally, this paper provides an evaluation of symbolic transformer model adapted via RegressLM and REMI tokenization, reporting distortion metrics and corresponding analyses. Finally, this paper analyzes the model's overall capabilities with handling various musical dimensions, relating these findings to perceptually annotated datasets and prior musical evaluation work.

Although regression algorithms and beat-aware tokenization measurably improve detection of severe and moderate errors, symbolic transformers are not yet sufficient for autonomous, nuanced evaluation without human oversight. Future work should pursue three directions:

1.Curriculum augmentation focused on timing distortions

2.Cross-dataset transfer and severity level analysis to better align model classifications with human judgements

3.Hybrid audio-symbolic models that handle nuance and explicit error labels

These steps will be crucial for applying transformer symbolic modeling to the task of performance assessment.

References

[1] W. Li, et al. Analysis of Piano Performance Characteristics by Deep Learning. Frontiers in Psychology, vol. 12, (2022)https: //www.frontiersin.org/articles/10.3389/fpsyg.2021.751406/full.

[2] C. Hawthorne, et al. Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset. Proceedings of ICLR, arXiv: 1810.12247, (2019)https: //arxiv.org/abs/1810.12247.

[3] J. Gardner, et al. MT3: Multi-Task Multitrack Music Transcription. arXiv, (2021)https: //arxiv.org/abs/2111.03017.

[4] Y. S. Huang, and Y. H. Yang. Pop Music Transformer: Beat-based Modeling and Generation of Expressive Pop Piano Compositions. Proceedings of the 28th ACM International Conference on Multimedia (MM '20), pp. 1180–1188, (2020)https: //arxiv.org/abs/2002.00212

[5] B. S. H. Chou, et al. Detecting Music Performance Errors with Transformers (Polytune). arXiv, (2025) https: //arxiv.org/abs/2501.02030.

[6] Q. Q. Kong, et al. GiantMIDI-Piano: A Large-Scale MIDI Dataset for Classical Piano Music. Transactions of the International Society for Music Information Retrieval, vol. 5, no. 1, pp. 49–58, (2022) https: //transactions.ismir.net/articles/10.5334/tismir.80

[7] J. Melechovský, R. Abhinaba, and H. Dorien. MidiCaps: A Large-Scale MIDI Dataset with Text Captions. arXiv, (2024) https: //arxiv.org/abs/2406.02255.

[8] J. Park, et al. Piano Performance Evaluation Dataset with Multi-Level Perceptual Features. Scientific Reports, vol. 14, (2024) https: //www.ncbi.nlm.nih.gov/pmc/articles/PMC11450231/.

[9] X. Q. Li, et al. Siamese Residual Neural Network for Musical Shape Evaluation in Piano Performance Assessment. arXiv, (2024) https: //arxiv.org/abs/2401.02566.

[10] Google DeepMind. google-deepmind/regress-lm. GitHub, https: //github.com/google-deepmind/regress-lm. (2025)

[11] X. W. Qu, et al. MuPT: A Generative Symbolic Music Pretrained Transformer. arXiv, (2024)https: //arxiv.org/abs/2404.06393.

[12] Y. J. Yan and Z. Y. Duan. Scoring Time Intervals using Non-Hierarchical Transformer for Automatic Piano Transcription. ISMIR arXiv, (2024) https: //arxiv.org/abs/2404.09466.

[13] Y. Akhauri, et al. Performance Prediction for Large Systems via Text-to-Text Regression.arXiv, (2025) https: //arxiv.org/abs/2506.21718.