Generating Piano Music with Transformers: A Comparative Study of Scale, Data, and Metrics

Jonathan Lehmkuhl* **Ábel Ilyés-Kun*** **Nico Bremes*** **Cemhan Kaan Özaltan***

Frederik Muthers* **Jiayi Yuan†**

Abstract

Although a variety of transformers have been proposed for symbolic music generation in recent years, there is still little comprehensive study on how specific design choices affect the quality of the generated music. In this work, we systematically compare different datasets, model architectures, model sizes, and training strategies for the task of symbolic piano music generation. To support model development and evaluation, we examine a range of quantitative metrics and analyze how well they correlate with human judgment collected through listening studies. Our best-performing model, a 950M-parameter transformer trained on 80K MIDI files from diverse genres, produces outputs that are often rated as human-composed in a Turing-style listening survey.

1 Introduction

Music exhibits hierarchical structure across multiple timescales, making transformer models with self-attention well-suited for modeling musical dependencies. Recently proposed transformer models can generate expressive sequences that capture both local patterns and global structure of MIDI data [1, 2]. A variety of MIDI datasets [3, 4, 5], tokenization schemes [6, 2, 7], transformer-based architectures [8, 9, 10], and evaluation strategies [11, 12] have been proposed in prior work on symbolic music generation. However, due to the lack of a unified evaluation framework for generated music, it remains an open challenge to determine how specific design choices in custom architectures, e.g., the tokenizer, embedding function, or attention mechanism, affect the quality of the generated music. While automatically computable metrics such as perplexity and musically informed objective metrics [11] are commonly used, their relationship to human perception of musical quality is not yet fully understood. We aim to address these challenges by presenting an empirical study on piano music generation using transformer models.

2 Methods

Models and training setups Our methodology is structured around five targeted experiments. First, we train scaled-down versions (62M, 155M, 439M, and 950M) of the Mistral 7B architecture [13] with sliding window attention and rotary position embedding on the MAESTRO dataset [3]. Furthermore, we pre-train models on a subset of the Aria-MIDI dataset [5] and compare them to models trained only on MAESTRO. We fine-tune a model pre-trained on Aria-MIDI on MAESTRO and compare it to a model trained from scratch only on MAESTRO to assess the effect of transfer learning. To guide generation, we prepend genre labels during training and examine their impact. A genre-conditioned 950M-parameter model is further evaluated in a Turing-like listening study. Finally, we fine-tune

*Equal contribution. RWTH Aachen University. Email: {jonathan.lehmkuhl, abel.ilyes-kun, nico.bremes, kaan.oezaltan, frederik.muthers}@rwth-aachen.de.

†University of Washington. Email: jiayiy9@cs.washington.edu.

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

the Moonbeam foundation model [10] on MAESTRO to provide a high-level benchmark for our custom Mistral-based models. Unlike standard lookup-based embeddings, Moonbeam uses a trainable sinusoidal embedding function [14] that enforces translational invariance and pitch transposability, claimed to yield more musically meaningful representations and improved test perplexity.

Tokenization and data preprocessing For tokenization, we use the REMI scheme [2], a widely adopted method that imposes metrical structure through position and bar tokens. Our REMI vocabulary contains 485 base tokens, which we extend using byte-pair encoding to learn a 30K-token vocabulary for more efficient representation. To balance efficiency and expressiveness, we fix the model context length to 1024 tokens, typically corresponding to 1–2 minutes of music. All MIDI files in our datasets are split into chunks of roughly 1024 tokens, and we generate samples of the same length during inference. Moonbeam compounds multiple events into single note-level tokens, producing sequences three times shorter than REMI. For a fair comparison, we adjust Moonbeam's context length to match the effective amount of training data seen by the Mistral-based models.

Datasets and data augmentation Various datasets of piano music in MIDI format are now available for symbolic music research [3, 15, 16, 17, 18, 5]. For our experiments, we use the MAESTRO dataset [3], containing 200 hours of professional classical piano, and a curated subset of 100K files from Aria-MIDI [5], which spans multiple genres of automatically transcribed piano recordings. To improve generalization and robustness, we augment the data by randomly altering pitch, velocity, note duration, and tempo within perceptually plausible ranges.

Subjective evaluation Subjective evaluation is our target metric, as our goal is to generate music perceived as enjoyable, original, and human-like. We conduct listening tests in which five participants rate samples on a five-point Likert scale along three dimensions: **pleasingness** (how enjoyable the music is), **authenticity** (how natural or human-like it sounds), and **novelty** (how original or unique it feels) [19], with samples presented in randomized order to avoid bias. Inspired by the original Turing Test [20], we also conduct a musical Turing-like test, asking listeners to classify excerpts as human-composed, AI-generated, or uncertain.

Objective evaluation We use **perplexity** (PPL) to assess token-level fit, though it does not capture higher-level musical structure. To evaluate global patterns, we employ distributional metrics comparing feature-level distributions of generated and training samples. **Fréchet music distance** (FMD) [12] measures distances between multivariate Gaussian embeddings extracted from MIDI files. We also compute mean and standard deviation of musical features (pitch count, range, intervals, note count, inter-onset interval) and compare datasets via histograms of intra-set and inter-set Euclidean distances, smoothed with kernel density estimation. Similarity is quantified using **Kullback–Leibler divergence** (KLD) [21] and **overlapping area** (OA), with low KLD and high OA indicating strong resemblance to the training data. We compute the above metrics at intermediate checkpoints during training.

3 Experiments and results

We provide detailed results as supplementary material in the appendix: Training curves are shown in Figures 1–6, the results of the musical Turing-like test are in Table 3 and Figures 7–12, and Tables 4–6 present the complete results of the objective and subjective metrics. Our code[3] is available online.

3.1 Different models on MAESTRO

We trained four Mistral-based models of varying sizes on the MAESTRO dataset for 225 epochs each. Subjective evaluation of generated samples reveals that, despite overfitting, generation quality generally improves with model size. For overfitting models, later checkpoints often produce more musically coherent outputs, even as validation loss worsens. While this might suggest memorization, our listening impressions indicate otherwise: the samples, though more structured and stylistically consistent, remain too flawed to be mistaken for exact copies of training pieces. This points to a fundamental tradeoff when training large models on small datasets like MAESTRO, especially for

[3] https://github.com/kaanozaltan/piano-transformer

creative domains. To learn high-level musical features, the models must overfit to some extent. While this limits generalization, it improves alignment with the training distribution—our primary goal in this setting. Table 1 presents subjective evaluation results alongside PPL, FMD, KLD, and OA for the final checkpoints. FMD, KLD, and OA closely align with human judgments in ranking model quality, whereas PPL diverges for the two largest models—supporting the view that token-level metrics are insufficient for evaluating musical plausibility. Interestingly, the 950M model underperforms the 439M model not only on KLD and OA, but also in the subjective evaluation at the final checkpoint.

For fine-tuning, we used the two available Moonbeam model checkpoints with 309M and 839M parameters, both pre-trained on a collection of datasets worth over 80K hours of music. We fine-tuned both models for 150 epochs on MAESTRO. We selected the 839M fine-tuned model and a context size of 512 for comparison with our custom models. We calculated the objective metrics on the same reference set as for our custom models with a fidelity of 1000 samples at the checkpoint corresponding to step 3000. Table 1 shows the comparison of the objective metrics and the scores of the subjective listening test across the models. The 439M and 950M custom models outperform Moonbeam on the subjective listening test.

Table 1: Subjective and objective evaluation of the custom models and the Moonbeam model trained on MAESTRO. Mean S. column indicates mean subjective evaluation scores. MB indicates the Moonbeam model. PPL not computed for Moonbeam.

Model	Mean S.	Pleas.	Auth.	Nov.	FMD ↓	KLD ↓	OA ↑	PPL ↓
62M	2.84	2.85	2.78	2.89	210.01	0.36	62.85%	53.65
155M	2.86	2.89	2.66	3.03	187.53	0.38	68.46%	21.32
439M	**3.22**	**3.30**	**3.02**	**3.35**	**176.40**	**0.25**	71.60%	4.82
950M	3.17	3.27	2.98	3.27	176.80	0.29	68.33%	**2.78**
MB 839M	2.91	2.90	2.74	3.08	194.70	0.51	**82.21%**	–

3.2 Fine-tuning on MAESTRO

While the Aria dataset provides a large and diverse collection of MIDI files, it is automatically transcribed from internet audio, which can introduce noise and errors. MAESTRO, in contrast, contains high-quality recordings by professional pianists, offering more reliable musical detail. To leverage Aria's scale while benefiting from MAESTRO's quality, we applied transfer learning by fine-tuning models pre-trained on the multi-genre subset of Aria-Deduped. We compared partial fine-tuning (155M-F-P), freezing the first eight blocks to retain general musical structure, with full fine-tuning (155M-F-F), updating all parameters to fully adapt to MAESTRO.

As shown in Table 2, both variants outperform the MAESTRO-only baseline across most metrics. Full fine-tuning achieves the best overall performance, suggesting that allowing all layers to adapt is beneficial, likely due to the genre gap between the multi-genre pre-training data and classical MAESTRO. The underperformance of partial fine-tuning indicates that low-level musical representations differ across genres and benefit from complete adaptation.

Table 2: Subjective and objective evaluation of models pre-trained on Aria-Deduped and fine-tuned on MAESTRO compared with the 155M MAESTRO model. Mean S. column indicates mean subjective evaluation scores.

Model	Mean S.	Pleas.	Auth.	Nov.	FMD ↓	KLD ↓	OA ↑
155M	2.85	2.80	2.80	2.95	187.53	0.38	68.46%
155M-F-P	3.09	3.13	2.95	3.20	193.45	0.34	70.93%
155M-F-F	**3.25**	**3.30**	**3.10**	**3.35**	**187.34**	**0.30**	**72.39%**

3.3 Integrating genre information

The Aria dataset contains multiple musical genres with differing distributional properties. To better control output style, we trained a model conditioned on genre tokens. Each file's genre was extracted from the metadata, embedded into the MIDI, and prepended as a token to the sequence. During

generation, the model could be conditioned by providing only the genre token. We trained a 155M-parameter model on 35K MIDI files from the same Aria subset used previously, allowing direct comparison with a model trained without genre tokens. Results show only minor differences in objective metrics, suggesting that genre tokens alone do not significantly improve the model's understanding of the data. Subjective listening, however, indicated that genre-conditioned samples align well with the intended style, with the 950M model producing particularly high-quality outputs where conditioned genres were clearly apparent.

3.4 Musical Turing-like test

Since the 950M model with genre conditioning produced the highest-quality samples, we conducted a musical Turing-like test using sequences from the five most prevalent Aria-Deduped genres. For each genre, five generated and five human-composed sequences were selected from a pool of thirty samples. Excluding "unsure" responses, participants achieved 61.2% accuracy; counting "unsure" as incorrect reduced it to 53.6%. Precision, recall, and F1-scores were similar for human and generated samples, and the confusion matrix showed roughly a third of samples misclassified, indicating many generated sequences were perceptually similar to human music. Accuracy by genre ranged from 48% (soundtrack) to 62% (pop), with biases: classical and jazz leaned "generated," while pop, rock, and soundtrack leaned "human." Overall, responses were 44.4% human, 43.2% generated, and 12.4% unsure, showing slight bias. Unsure rates were slightly higher for generated sequences (14.4% vs. 10.4%), suggesting confusion remained high across genres.

4 Conclusion

This study systematically compared transformer models for unconditional piano music generation, varying model scale, dataset size, pre-training and fine-tuning strategies, conditioning methods, and architectures. Larger models generally improved subjective quality, though excessive overfitting on the small MAESTRO dataset sometimes reduced performance. Objective metrics FMD, KLD, and OA aligned reasonably well with human judgments, whereas PPL consistently improved with model size but did not always reflect perceived quality. Pre-training on the larger Aria-MIDI dataset enhanced generation quality and mitigated overfitting, with fine-tuning on MAESTRO further improving both subjective and objective results. Genre conditioning allowed style-specific control, and the largest conditioned model frequently produced outputs rated as human-composed in a Turing-style listening test.

Training larger models on MAESTRO revealed a trade-off between overfitting and musical coherence. Overfitting sometimes improved alignment with the training distribution, producing more coherent outputs without simple memorization, but excessive overfitting could reduce quality. On larger datasets such as Aria-MIDI, this trade-off was less pronounced. While FMD, KLD, and OA generally correlated with subjective judgments, no metric fully captured musical quality. Dataset scale and transfer learning were critical, with pre-training on Aria-MIDI improving generalization and enabling MAESTRO fine-tuning to surpass MAESTRO-only models.

While Moonbeam's domain-specific embedding space offers theoretical advantages, our custom models performed competitively despite relying on conventional lookup embeddings. This raises questions about the practical benefits of such specialized architectures and whether specific factors may have limited the effectiveness of fine-tuning Moonbeam on MAESTRO.

Future work should further explore the trade-off between overfitting and generation quality, particularly using high-level musical metrics rather than cross-entropy loss, to better understand when overfitting begins to harm outputs. The benefits of specialized architectures for music generation remain an open question. Future work could probe the features captured at different layers of the model and assess whether the embedding space preserves pitch-transposition relationships in deeper layers. This could be evaluated by measuring reconstruction loss on transposed inputs at various stages in the network. The establishment of robust benchmarks for fair comparison between models from different studies would be desirable. Standardized evaluation frameworks would greatly improve the comparability and reproducibility of results in this field.

References

[1] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Ian Simon, Curtis Hawthorne, Noam Shazeer, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. Music transformer: Generating music with long-term structure. In *International Conference on Learning Representations*, 2019.

[2] Yu-Siang Huang and Yi-Hsuan Yang. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1180–1188, 2020.

[3] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International Conference on Learning Representations*, 2019.

[4] Colin Raffel. *Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching*. Columbia University, 2016.

[5] Louis Bradshaw and Simon Colton. Aria-midi: A dataset of piano midi files for symbolic music modeling. In *International Conference on Learning Representations*, 2025.

[6] Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan. This time with feeling: Learning expressive musical performance. *Neural Computing and Applications*, 32(4):955–967, 2020.

[7] Dimitri von Rütte, Luca Biggio, Yannic Kilcher, and Thomas Hofmann. Figaro: Controllable music generation using learned and expert features. In *The Eleventh International Conference on Learning Representations*, 2023.

[8] Botao Yu, Peiling Lu, Rui Wang, Wei Hu, Xu Tan, Wei Ye, Shikun Zhang, Tao Qin, and Tie-Yan Liu. Museformer: Transformer with fine-and coarse-grained attention for music generation. *Advances in neural information processing systems*, 35:1376–1388, 2022.

[9] Cong Jin, Tao Wang, Xiaobing Li, Chu Jie Jiessie Tie, Yun Tie, Shan Liu, Ming Yan, Yongzhi Li, Junxian Wang, and Shenze Huang. A transformer generative adversarial network for multi-track music generation. *CAAI Transactions on Intelligence Technology*, 7(3):369–380, 2022.

[10] Zixun Guo and Simon Dixon. Moonbeam: A midi foundation model using both absolute and relative music attributes. *arXiv preprint arXiv:2505.15559*, 2025.

[11] Li-Chia Yang and Alexander Lerch. On the evaluation of generative models in music. *Neural Computing and Applications*, 32(9):4773–4784, 2020.

[12] Jan Retkowski, Jakub Stępniak, and Mateusz Modrzejewski. Frechet music distance: A metric for generative symbolic music evaluation. *arXiv preprint arXiv:2412.07948*, 2024.

[13] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[14] Zixun Guo, Jaeyong Kang, and Dorien Herremans. A domain-knowledge-inspired music embedding space and a novel attention mechanism for symbolic music modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5070–5077, 2023.

[15] Drew Edwards, Simon Dixon, and Emmanouil Benetos. Pijama: Piano jazz with automatic midi annotations. *Transactions of the International Society for Music Information Retrieval*, 2023.

[16] Qiuqiang Kong, Bochen Li, Jitong Chen, and Yuxuan Wang. Giantmidi-piano: A large-scale midi dataset for classical piano music. *arXiv preprint arXiv:2010.07061*, 2020.

[17] Huan Zhang, Jingjing Tang, Syed Rifat Mahmud Rafee, and Simon Dixon György Fazekas. Atepp: A dataset of automatically transcribed expressive piano performance. In *ISMIR 2022 Hybrid Conference*, 2022.

[18] Lucas N. Ferreira, Levi H. S. Lelis, and Jim Whitehead. Computer-generated music for tabletop role-playing games. In *Proceedings of the Sixteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, AIIDE'20. AAAI Press, 2020.

[19] Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang. Midinet: A convolutional generative adversarial network for symbolic-domain music generation. *arXiv preprint arXiv:1703.10847*, 2017.

[20] Alan Mathison Turing. Computing machinery and intelligence. *Mind*, 49:433–460, 1950.

[21] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

A Training Curves

Figure 1: Training curves for the MAESTRO models of different sizes.

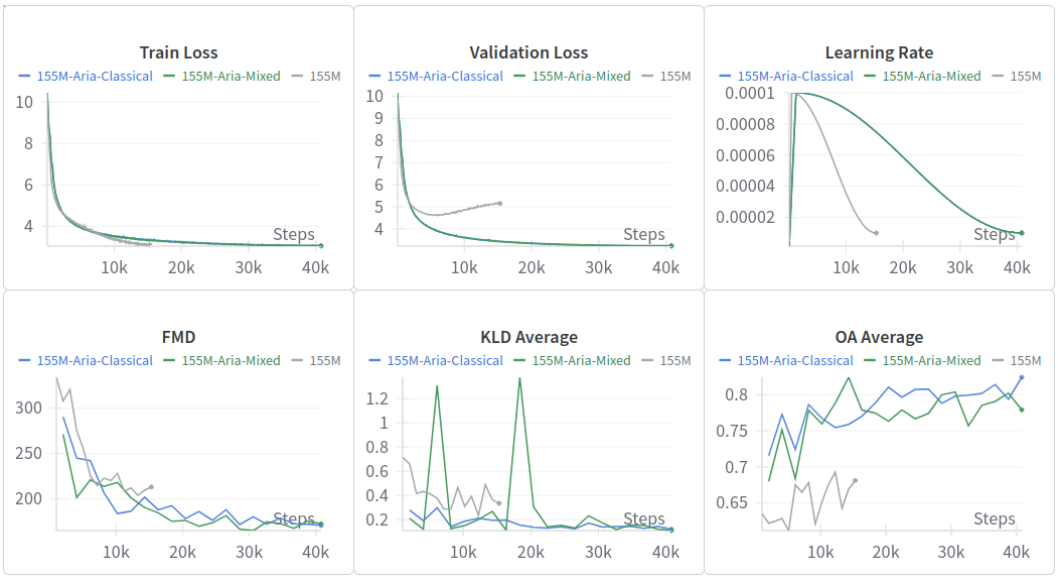

Figure 2: Training curves for the models pre-trained on Aria-Deduped compared to the 155M model trained only on MAESTRO.

Figure 3: Training curves for the models pre-trained on Aria-Deduped and fine-tuned on MAESTRO.

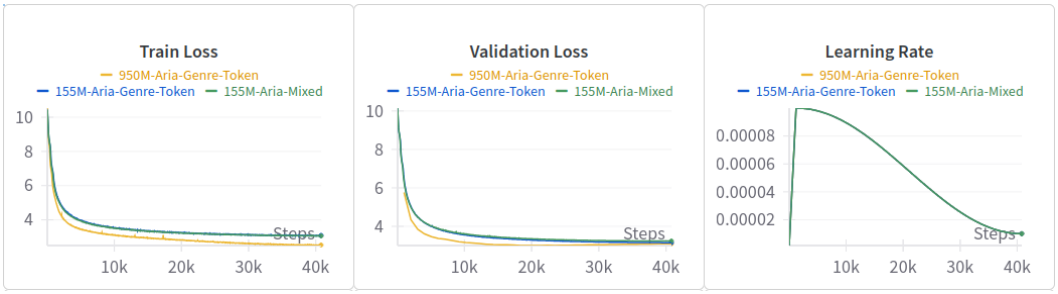

Figure 4: Training curves for the models with integrated genre information.

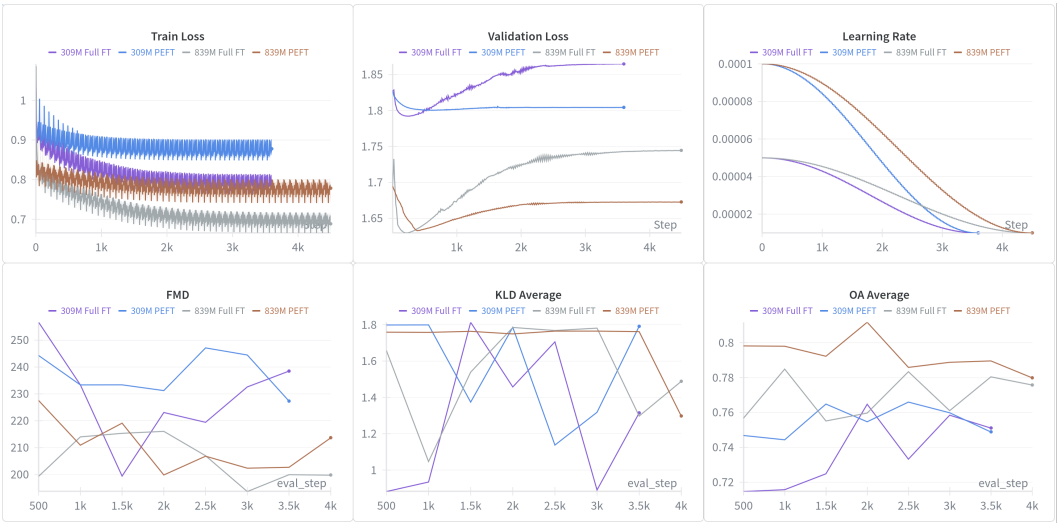

Figure 5: Fine-tuning curves for Moonbeam with context size 512.

Figure 6: Fine-tuning curves for Moonbeam with context size 1024.

B Musical Turing-like Test Analysis Results

Table 3: Precision, Recall, and F1-Scores for the musical Turing-like test. "Unsure" responses were considered as incorrect.

Class	Precision	Recall	F1-score
Human	**62.16%**	**55.20%**	**58.47%**
Generated	60.19%	52.00%	55.79%

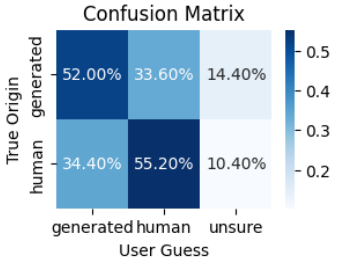

Figure 7: Confusion Matrix for the musical Turing-like test.

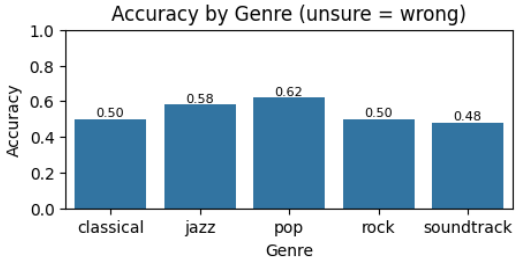

Figure 8: Accuracy by genre in the musical Turing-like test.

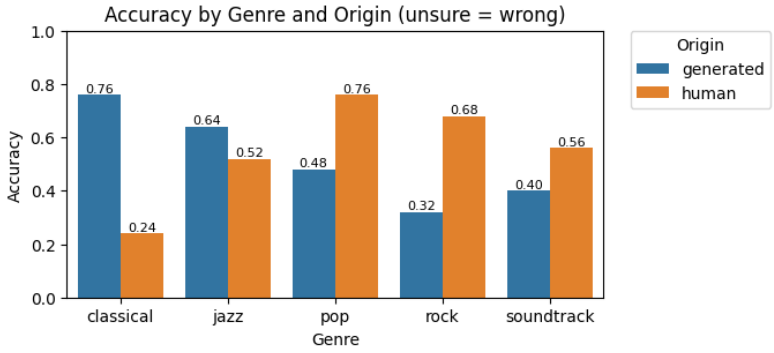

Figure 9: Accuracy by genre and origin in the musical Turing-like test.

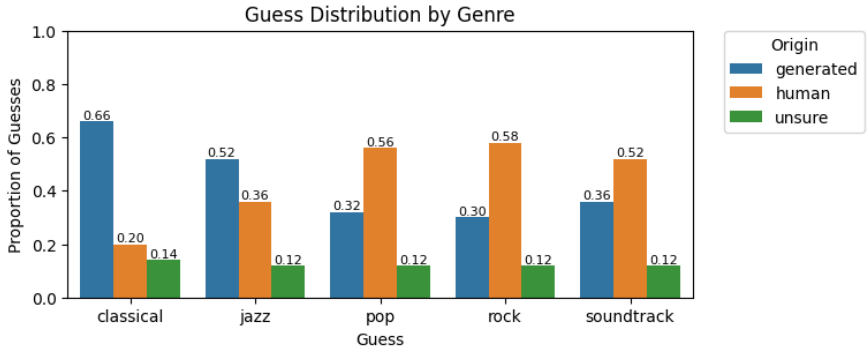

Figure 10: Distribution of participant guesses by genre in the musical Turing-like test.

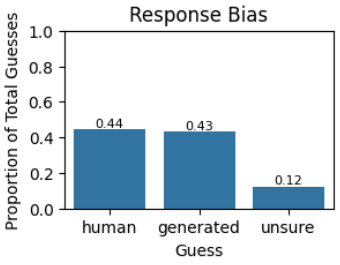

Figure 11: Overall response bias in the musical Turing-like test.

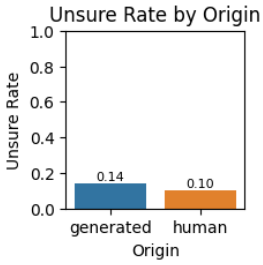

Figure 12: Unsure rate by origin in the musical Turing-like test.

C Full Evaluation Tables

Table 4: Subjective and objective evaluations of models trained on MAESTRO.

Model	**62M**		**155M**		**439M**		**950M**		**MAESTRO**	
Subjective Evaluation										
Pleasingness	2.85		2.89		3.30		3.27			
Authenticity	2.78		2.66		3.02		2.98			
Novelty	2.89		3.03		3.35		3.27			
Average	2.84		2.86		3.22		3.17			
Absolute Evaluation										
	Mean	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**
total_used_pitch	38.15	9.24	42.09	9.07	43.99	8.45	42.73	8.77	54.90	9.82
pitch_range	51.39	11.05	54.24	10.17	53.19	9.53	52.61	10.09	63.17	9.23
avg_pitch_shift	9.36	4.25	9.88	3.51	11.05	3.31	10.67	3.82	12.35	2.79
total_used_note	476.18	99.74	497.79	96.82	518.03	72.06	514.75	73.43	609.97	51.24
avg_IOI	0.15	0.10	0.13	0.09	0.14	0.10	0.14	0.09	0.09	0.05
Average Distance to MAESTRO	0.33	10.48	0.25	9.61	0.23	4.61	0.25	5.04	0.00	0.00
Relative Evaluation										
	KLD	**OA**	**KLD**	**OA**	**KLD**	**OA**	**KLD**	**OA**		
total_used_pitch	0.03	71.63%	0.03	81.79%	0.01	86.46%	0.03	82.64%		
total_pitch_class_hist	0.20	75.52%	0.61	80.20%	0.01	88.03%	0.02	86.38%		
pitch_range	0.01	78.90%	0.13	85.96%	0.01	83.63%	0.02	81.62%		
avg_pitch_shift	0.05	76.91%	0.00	86.25%	0.00	93.18%	0.01	87.74%		
total_used_note	1.74	22.36%	1.35	27.13%	1.20	30.02%	1.20	30.92%		
avg_IOI	0.03	77.42%	0.16	83.25%	0.09	81.00%	0.08	79.32%		
note_length_hist	0.49	45.21%	0.44	47.20%	0.41	50.66%	0.58	44.08%		
note_length_transition_matrix	0.30	54.81%	0.30	55.90%	0.26	59.83%	0.37	53.94%		
Average	0.36	62.85%	0.38	68.46%	0.25	71.60%	0.29	68.33%		
FMD	210.01		187.53		176.40		176.38			
Perplexity	53.65		21.32		4.82		2.78			

Table 5: Subjective and objective evaluation of models pre-trained on Aria-Deduped compared with the 155M MAESTRO model.

Model	**155M**		**155M-A-M**		**155M-A-C**		**MAESTRO**	
Subjective Evaluation								
Pleasingness	2.80				3.25			
Authenticity	2.80				3.03			
Novelty	2.95				3.25			
Average	2.85				3.18			
Absolute Evaluation								
	Mean	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**
total_used_pitch	42.09	9.07	30.11	8.64	32.39	10.67	54.90	9.82
pitch_range	54.24	10.17	48.93	11.54	50.34	12.01	63.17	9.23
avg_pitch_shift	9.88	3.51	9.71	3.02	10.18	3.32	12.35	2.79
total_used_note	497.79	96.82	436.02	96.08	431.72	130.27	609.97	51.24
avg_IOI	0.13	0.09	0.27	0.18	0.25	0.34	0.09	0.05
Average Distance to MAESTRO	0.25	9.61	0.65	9.74	0.58	16.70	0.00	0.00
Relative Evaluation								
	KLD	**OA**	**KLD**	**OA**	**KLD**	**OA**		
total_used_pitch	0.03	81.79%	0.66	52.09%	0.53	56.25%		
total_pitch_class_hist	0.61	80.20%	0.25	70.96%	0.29	68.59%		
pitch_range	0.13	85.96%	0.20	71.89%	0.19	73.49%		
avg_pitch_shift	0.00	86.25%	0.06	85.88%	0.03	89.43%		
total_used_note	1.35	27.13%	2.28	24.10%	2.65	15.64%		
avg_IOI	0.16	83.25%	0.65	51.63%	0.66	49.49%		
note_length_hist	0.44	47.20%	1.48	41.16%	1.21	40.00%		
note_length_transition_matrix	0.30	55.90%	0.97	50.29%	0.87	49.21%		
Average	0.38	68.46%	0.82	56.00%	0.80	55.26%		
FMD	187.53		292.46		249.84			

Table 6: Subjective and objective evaluation of models pre-trained on Aria-Deduped and fine-tuned on MAESTRO compared with the 155M MAESTRO model.

Model	**155M**		**155M-F-P**		**155M-F-F**		**MAESTRO**	
Subjective Evaluation								
Pleasingness	2.80		3.13		3.30			
Authenticity	2.80		2.95		3.10			
Novelty	2.95		3.20		3.35			
Average	2.85		3.09		3.25			
Absolute Evaluation								
	Mean	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**
total_used_pitch	42.09	9.07	41.78	8.89	41.41	8.08	54.90	9.82
pitch_range	54.24	10.17	55.63	9.71	56.61	10.09	63.17	9.23
avg_pitch_shift	9.88	3.51	11.30	3.56	10.48	3.59	12.35	2.79
total_used_note	497.79	96.82	514.01	77.59	521.21	57.95	609.97	51.24
avg_IOI	0.13	0.09	0.12	0.08	0.11	0.07	0.09	0.05
Average Distance to MAESTRO	0.25	9.61	0.19	5.71	0.18	2.03	0.00	0.00
Relative Evaluation								
	KLD	**OA**	**KLD**	**OA**	**KLD**	**OA**		
total_used_pitch	0.03	81.79%	0.09	80.45%	0.09	80.98%		
total_pitch_class_hist	0.61	80.20%	0.14	78.01%	0.17	75.66%		
pitch_range	0.13	85.96%	0.02	88.78%	0.02	89.58%		
avg_pitch_shift	0.00	86.25%	0.04	88.44%	0.06	84.66%		
total_used_note	1.35	27.13%	1.30	30.39%	1.12	32.05%		
avg_IOI	0.16	83.25%	0.02	90.27%	0.03	90.51%		
note_length_hist	0.44	47.20%	0.62	52.51%	0.50	59.45%		
note_length_transition_matrix	0.30	55.90%	0.46	60.06%	0.36	66.20%		
Average	0.38	68.46%	0.34	70.93%	0.30	72.39%		
FMD	187.53		193.45		187.34			