

1. Appendix

1.1. Training curves



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

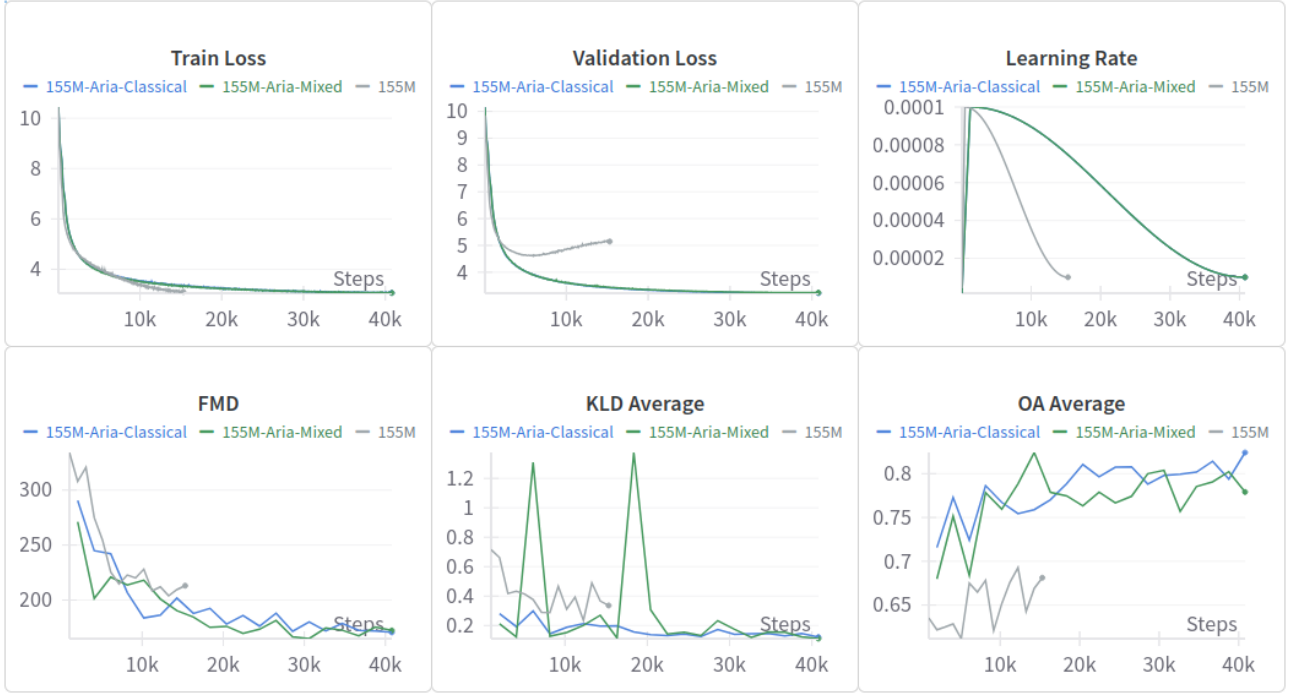


Figure 2. Training curves for the models pre-trained on Aria-Deduped compared to the 155M model trained only on MAESTRO.



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.

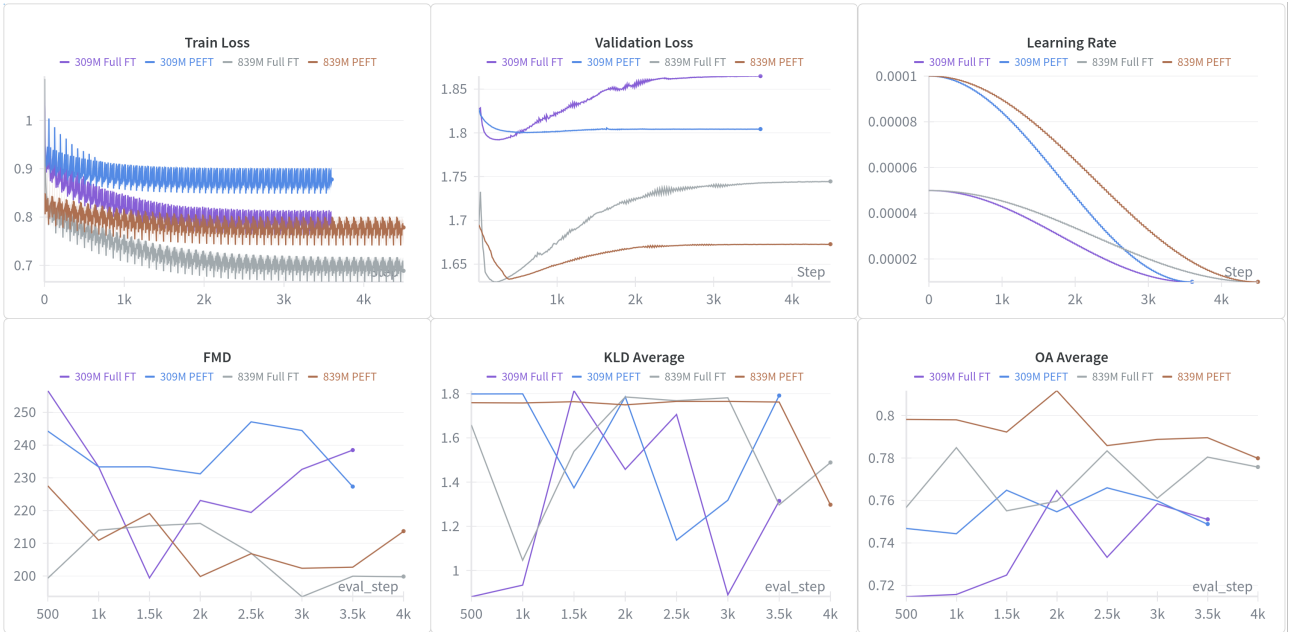


Figure 5. Fine-tuning curves for Moonbeam with context size 512.



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

1.2. Musical Turing-like rest analysis results

Table 8. 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%

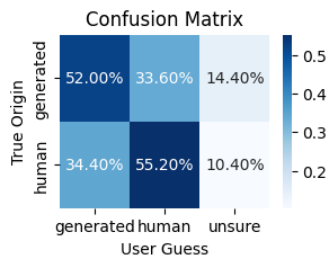


Figure 7. Confusion Matrix for the musical Turing-like test.

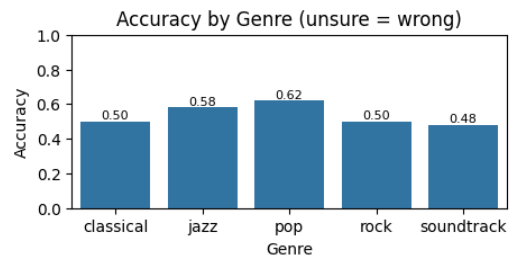


Figure 8. Accuracy by genre in the musical Turing-like test.

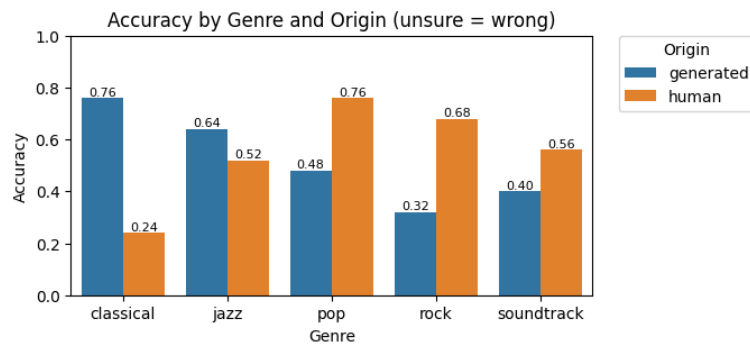


Figure 9. Accuracy by genre and origin in the musical Turing-like test.

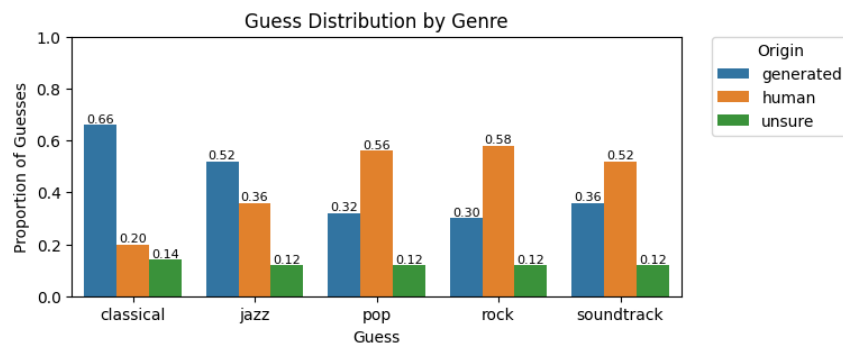


Figure 10. Distribution of participant guesses by genre in the musical Turing-like test.

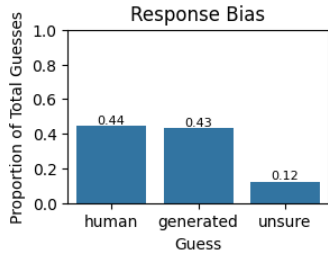


Figure 11. Overall response bias in the musical Turing-like test.

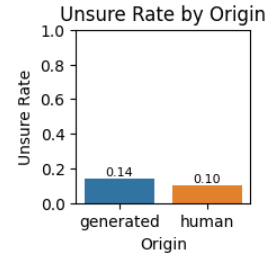


Figure 12. Unsure rate by origin in the musical Turing-like test.

2. Full Evaluation Tables

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 9. Subjective and objective evaluations of models trained on MAESTRO.

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 10. Subjective and objective evaluation of models pre-trained on Aria-Deduped 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			

Table 11. Subjective and objective evaluation of models pre-trained on Aria-Deduped and fine-tuned on MAESTRO compared with the 155M MAESTRO model.