MUPT: A GENERATIVE SYMBOLIC MUSIC PRETRAINED TRANSFORMER

Xingwei Qu*, Yuelin Bai*, Yinghao Ma*, Ziya Zhou, Ka Man Lo, Jiaheng Liu, Ruibin Yuan, Lejun Min, Xueling Liu, Tianyu Zhang, Xinrun Du, Shuyue Guo, Yiming Liang, Yizhi Li, Shangda Wu, Junting Zhou, Tianyu Zheng, Ziyang Ma, Fengze Han, Wei Xue, Gus Xia, Emmanouil Benetos, Xiang Yue, Chenghua Lin, Xu Tan, Stephen W. Huang, Jie Fu, Ge Zhang

MANCHESTER 1824 The University of Manchester

M-A-P

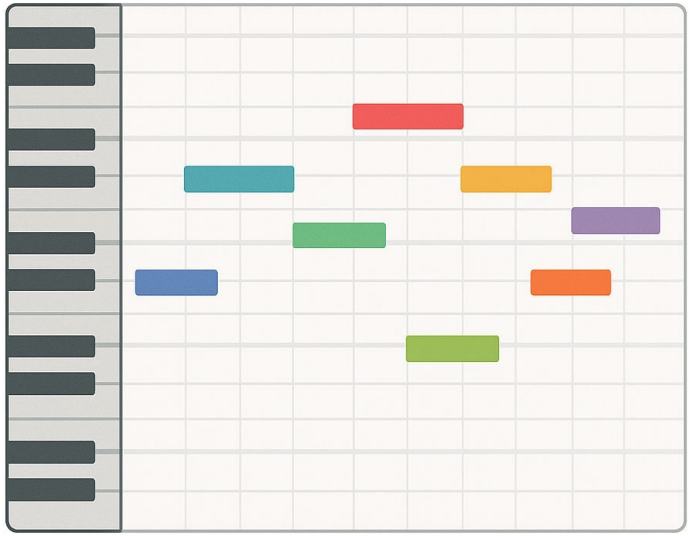
Queen Mary University of London

Motivation: To Develop A Symbolic Music GPT

- MuPT: Long-context LLMs (8,192 tokens) trained on ABC notation
- SMT-ABC tokeniser: A compact, structured representation that enhances coherence and preserves musical patterns using auto-regressive model l ing.
- SMS Law: Utilise performance on small models (200M to 1B parameters) to predict performance on large models (2B-4B) on limited data (~32B tokens)
- Open Source: release all models ' training checkpoints and training code (fixed some Megatron-LM issue s) to support open research in symbolic music generation.

MIDI vs. ABC Notation

MIDI



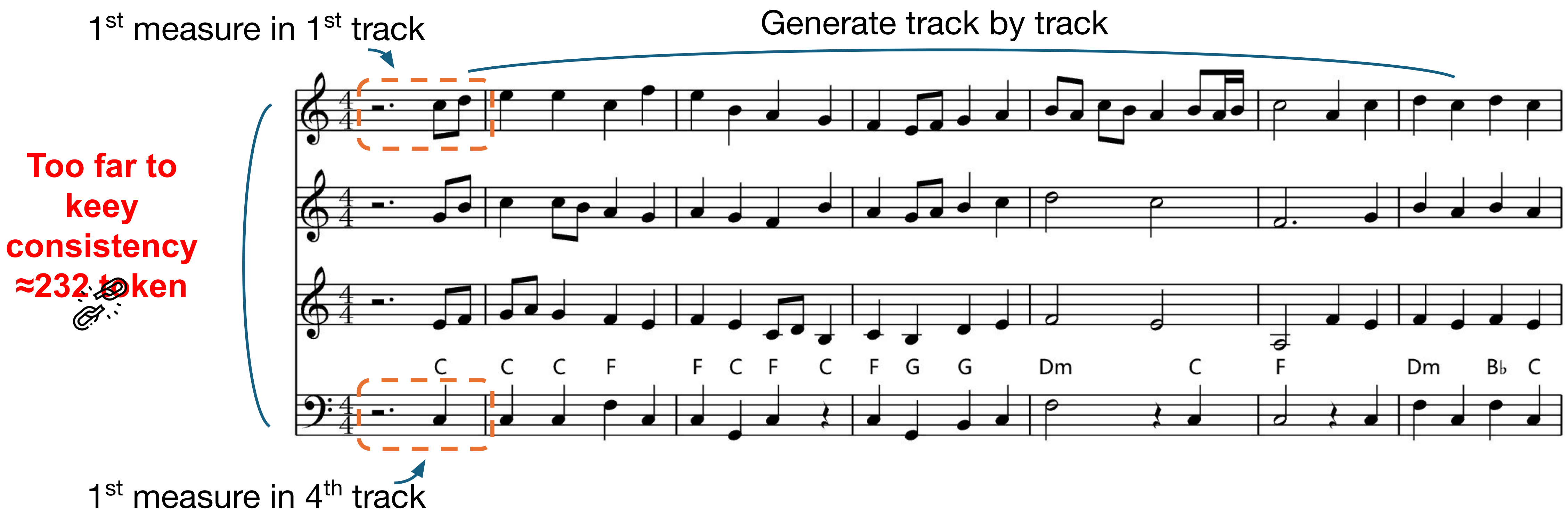
- Lack of structural coherence
- Difficulty in handling long sequences

ABC

- Better readability and clear structure
- Compact representation
- Less performance information

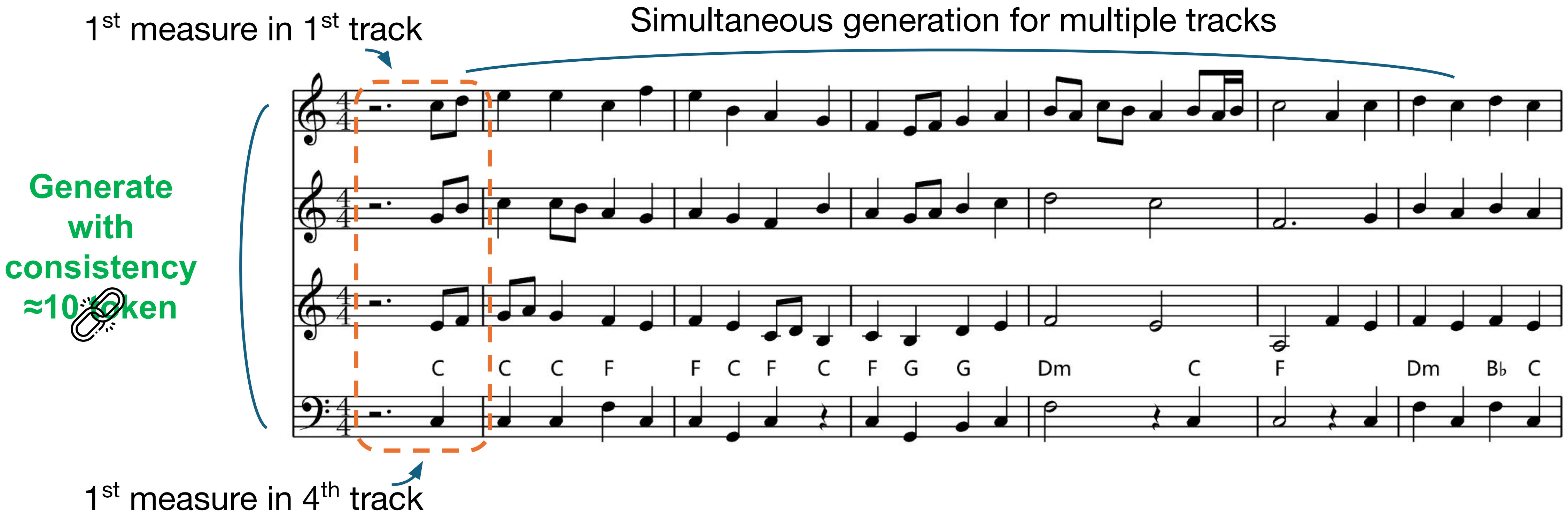
Disadvantage of Multi-Track ABC Notation

Obsevation: The first measure in the first track is too far away from the first measure in the following tracks, which can cause synchronization problems between the tracks.



Synchronized Multi-Track ABC (SMT-ABC) Notation

SMT-ABC: Generate multiple tracks for the same measure simultaneously



Comparison of Training Strategies

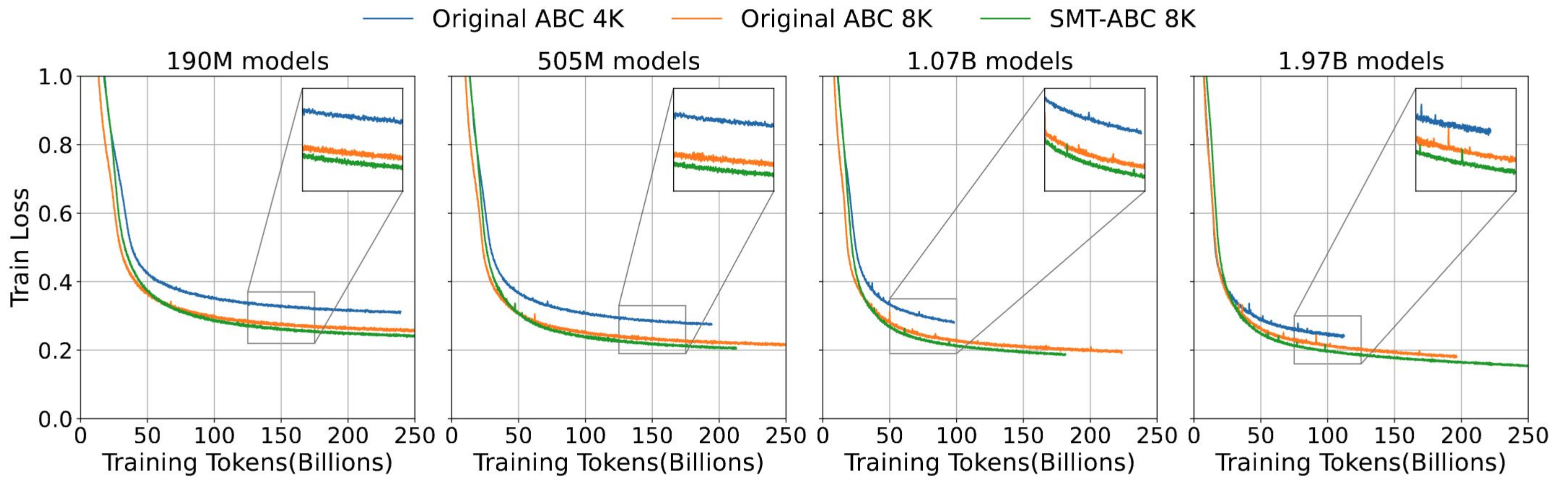
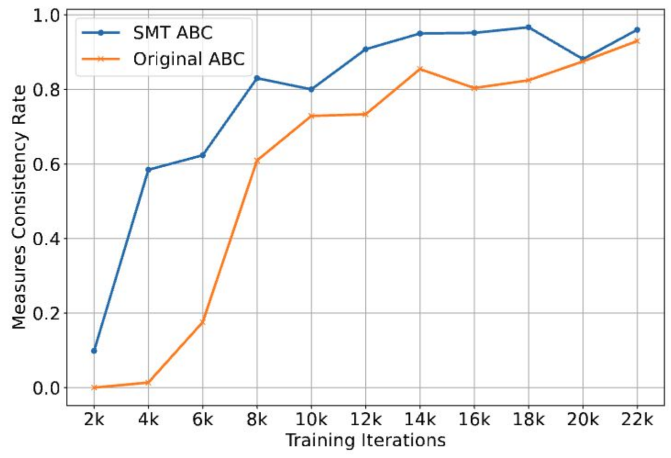


Figure 4: Training Loss for different model sizes and training strategy.



Measure consistency of SMT-ABC and ABC

- Longer context lengths help the model converge better, and our proposed SMT-ABC further accelerates the convergence speed.
- SMT-ABC model generates sequences with significantly higher consistency, promoting structural uniformity across tracks and enhancing the coherence and usability of compositions.

Constrained TrainingSet on Symbolic Music

- MuPT is trained on a large-scale, diverse symbolic music dataset that covers a wide range of music genres and track numbers.
- 32B tokens in total, including most of the publicly-available corpus, including MIDI-score2ABC. Far from enough for LLMs

Data Type	Count	Pct. (%)	Avg. Tks
Single Track	3.5M	51.2	450
2 Tracks	605K	8.7	2.0K
3 Tracks	412K	5.9	3.1K
4 Tracks	632K	9.0	4.2K
5 Tracks	362K	5.2	5.2K
6 Tracks	248K	3.6	6.7K
7 Tracks	176K	2.5	8.2K
8 Tracks	149K	2.1	10.1K
9 Tracks	104K	1.5	10.3K
10 Tracks	88K	1.3	11.8L
11+ Tracks	633K	9.1	25.9K
Total	6.9M	100.00	4.53

Genre	Number of Songs
Pop	256k
Jazz	107k
Country	49k
Rock	217k
Disco	6k
World Music (including Latin)	47k
Folk	118k
R&B, Funk & Soul	63k
Classical	466k

Exploring the Scaling Laws of Symbolic Music Generation

- Predicting LLM performance (CrossEntropy loss) L on a valid set based on Scaling Law based on a number of parameters N and training data volume.
- Baseline fitting: Chinchilla Law (below)

$$L(N, D) = \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} + E$$

$$\arg\min_{N,D} L(N, D) \quad \text{s.t.} \quad \text{FLOPs}(N, D) = C$$

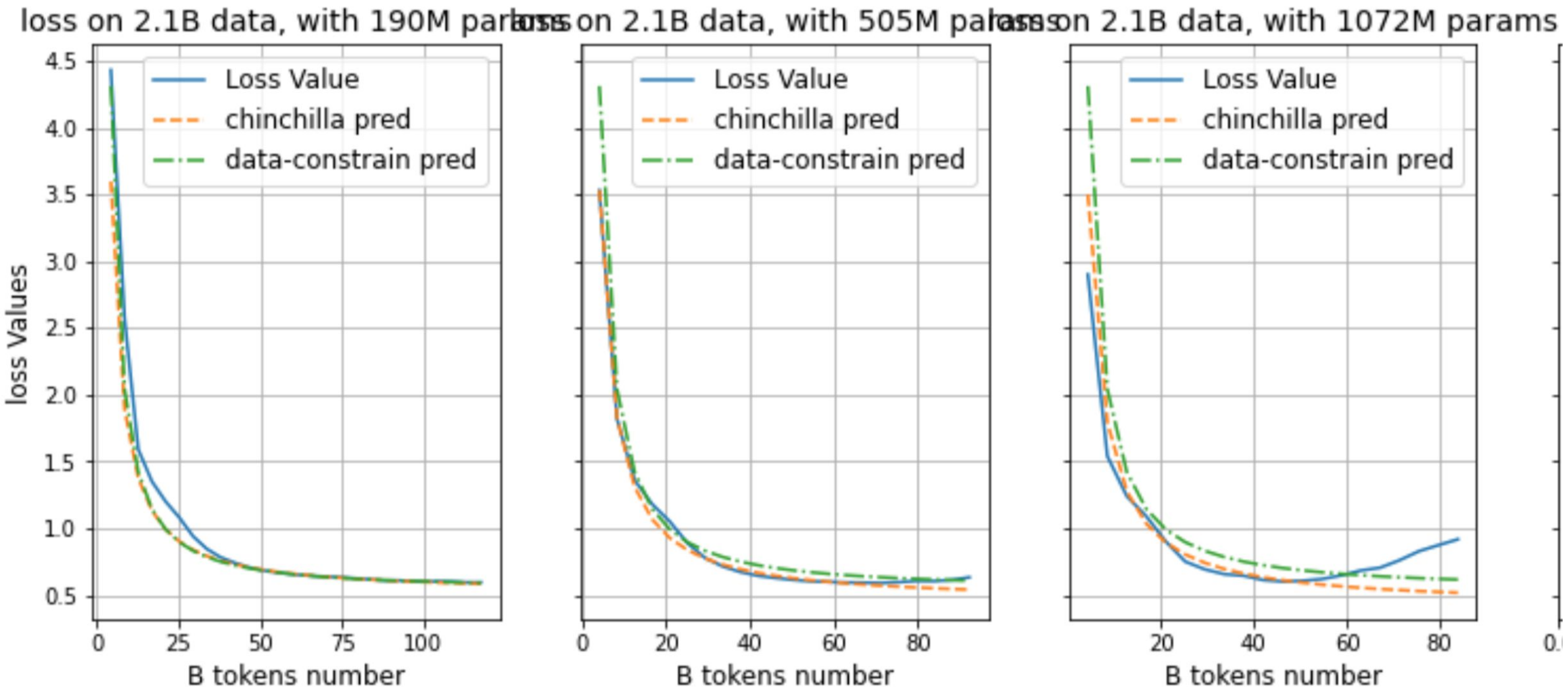
* Adaptation on limited data: Data-constrained Law (NeurIPS2023 best paper)

$$L(N, D, U_D) = \frac{A}{N'^{\alpha}} + \frac{B}{D'^{\beta}} + E$$

$$N' = U_N + U_N R_N^{\star}\left(1 - \exp\left(\frac{-R_N}{R_N^{\star}}\right)\right)$$

$$D' = U_D + U_D R_D^{\star}\left(1 - \exp\left(\frac{-R_D}{R_D^{\star}}\right)\right)$$

Failure of the baseline Scaling Law

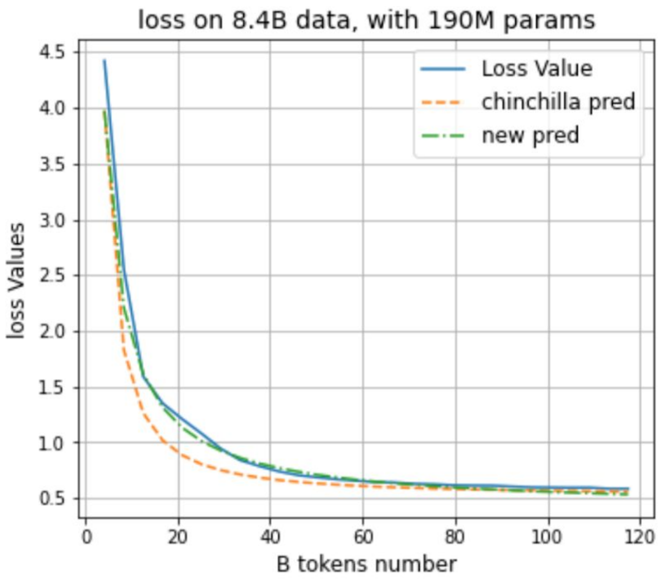


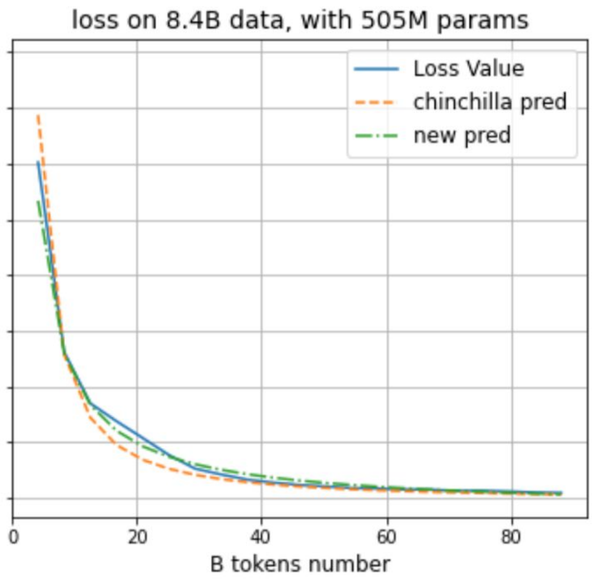
- Not fit when D is really small -> ND is not separable in the formula
- Overfitting after 2-3 epochs

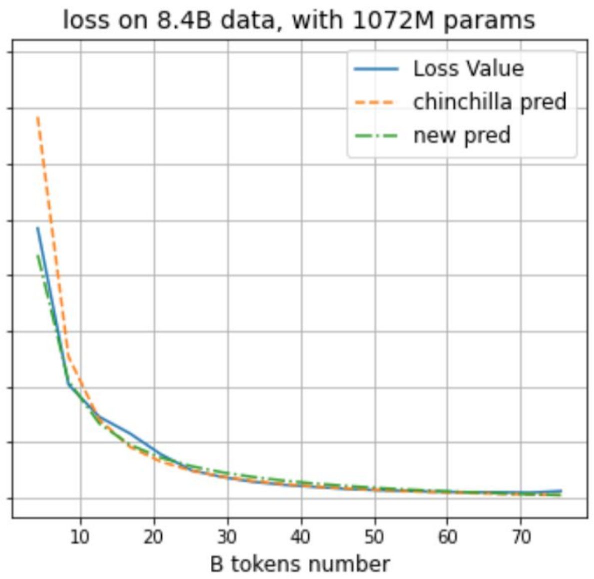
Symbolic Music Scaling (SMS) Law

$$L(N, D, U_D) = \frac{d}{N^{\alpha} \cdot D''^{\beta}} + \frac{A}{N^{\alpha}} + \frac{B}{D''^{\beta}} + E + L_{overfit}$$

$$L_{overfit} = \mathrm{GELU}\left\{k_d \cdot D + k_n \cdot \log(N) - k_u \cdot \log(U_D) - k_{in}\right\}$$







Formula Fitting Results of SMS Law

- Experiments: training on 2.1B tokens, 8.4B tokens, and 33.6B tokens repeatedly with 190M, 500M and 1.07B.
- Calculating R2 and Huber loss between authentic cross - entropy and predicted cross-entropy.
- Prediction: scaling to 1.97B can be better than larger (e.g. 4.23B)

Paramatic fit	R^2 Value (train) ↑	Huber Loss (train) ↓	R^2 Value (test) ↑	Huber Loss (test) ↓
Chinchilla law	0.9347	0.0109	-0.0933	0.0080
Data-Constrained law	0.7179	0.0206	0.1524	0.0071
Equation 11	0.9075	0.0129	0.3114	0.0073
Equation 2	0.9759	0.0102	0.8580	0.0062
SMS Law	**0.9780**	**0.0085**	**0.9612**	**0.0028**

Table 4: Comparison of parametric fitting performance of different scaling laws.

Music Elements Evaluation

- MuPT model outperforms both MIDI-based models and the ABC-based SOTA models on continuous generation including chatMusician.
- MuPT supports multi-track music generation (upper table), a feature missing in ChatMusician, making it more suited for realistic settings

System	PE	SC (%)	GC (%)
GT	2.708	96.80	93.46
MuPT-SMT	2.631	**97.48**	**93.45**
MuPT-Ori.	2.621	98.09	93.36
MMT	2.784	95.64	91.65
GPT-4	**2.783**	97.90	95.32
GT(st.)	2.617	98.39	93.25
MuPT-SMT(st.)	2.612	**98.20**	**93.39**
MuPT-Ori.(st.)	**2.619**	98.16	93.49
ChatMusician(st.)	2.664	98.55	94.47
MMT(st.)	2.808	95.88	91.60
GPT-4(st.)	2.686	99.27	95.72

Pitch entropy (PE)
Scale consistency (SC): counting the fraction of tones that were part of a standard scale and reporting the number for the best matching.
Groove consistency (GC): Rhythm metrics

A closer value to the ground truth (GT) is considered better.

Music Structure Evaluation

- MuPT surpassed GPT-4 by 17% and ChatMusician by 6% in terms of Intra Similarity and Repetition Rate, demonstrating its superior capability in handling complex musical compositions.

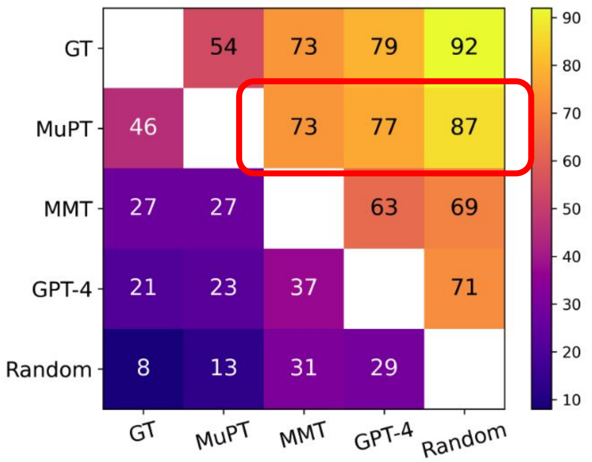
System	ITS	RR (%)
GT	0.3729	43.5
MuPT-SMT	**0.4193**	**43.7**
MMT	0.1767	-
GPT-4	0.3614	16.9
GT(st.)	0.4753	59.2
MuPT-SMT(st.)	**0.4507**	**52.6**
ChatMusician(st.)	0.5260	40.1
MMT(st.)	0.2158	-
GPT-4(st.)	0.4235	23.0

Intra-texture similarity (ITS): Average of self-deep-similarity in texture.
Repetition rate (RR): Percentage of sign ": |" appears in a generated set.

Both metrics evaluate the music structure.

Human Evaluation

- Subjective evaluations further validated MuPT's superiority, with over 70% preference ratings against both MMT and GPT-4, underscoring its appeal to human listeners.



Model A	Model B	Wins (A/B)	p-value
Human Works	MuPT	81/69	0.4237
	MMT	109/41	4.2249×10^{-6}
	GPT-4	119/31	6.6315×10^{-9}
	Random	138/12	4.4648×10^{-17}
MuPT	MMT	110/40	4.2249×10^{-6}
	GPT-4	115/35	6.6641×10^{-8}
	Random	131/19	1.3618×10^{-13}
MMT	GPT-4	95/55	0.0093
	Random	103/47	0.0001
GPT-4	Random	106/44	2.6691×10^{-5}

Table 7: Human evaluation of paired completions of musical excerpts generated by different sources given the first bar as the condition. The left is the matrix based on the AB test. Each row indicates the % of times listeners preferred instrumentals from that system compared to those from each system individually (N = 150). Ground truth is denoted by GT. i.e.77 means that listeners preferred MuPT over GPT-4 in 77% of cases. The right is the absolute win numbers and the corresponding p-value of each pair. P-values are reported by a Wilcoxon signed rank test.

Demos

- https://x.com/GeZhang86038849/status/1778620860737417491/video/2