# AGOGIC: AUDITING MUSIC TOKENIZATION UNDER MIDI TIME-UNIT SENSITIVITY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

CLaMP-2 Fréchet Music Distance (FMD) is used to rank text-to-MIDI generators, but its MIDI input encodes time in arbitrary file units. Multiplying both ticks per quarter and every delta tick by two or four preserves all note times, tempo messages and non-time events in 500 real scores, yet gives FMD 35.4/78.1 between copies of the same music. We retrain seven serializers on identical 84,576 MIDI–caption slices, 80-token prefixes and Qwen3.5-0.8B settings, evaluating one fixed 500-piece reference. The raw Beat-TSD-minus-PMT FMD gap of 136.4 contracts to 14.4 under a message-preserving common 220-TPQ export with at most 1.14 ms timing change. An exact common-unit export also changes the margins. We provide a note-only diagnostic that rounds absolute times to 1 ms and writes fixed units: all 500 real and 693 valid generated exact-rescale pairs become byte-identical. Canonical FMD for PMT/Beat-TSD/PerTok/REMI/Structured is 151.6/189.4/131.1/143.0/126.3; six-comparison sampling intervals on fixed seed-0 checkpoints place Beat-TSD above PMT, PerTok and Structured below it, and leave REMI unresolved. A within-source caption shuffle yields only a 0.015 instrument-F1 difference. The note-only remedy restores invariance to the tested exact tick rescaling; neither FMD protocol is validated against listening, and prompt-level control remains unresolved.

## 1 INTRODUCTION

Text-to-symbolic-music models turn a caption into an editable score with pitches, onsets, durations, dynamics and instruments. Before choosing a backbone, a system must choose a score serialization. Performance-event streams place fine time shifts among note events (Oore et al., 2018; Huang et al., 2019); beat-position streams align notes to a metrical grid (Huang & Yang, 2020; Fradet et al., 2021); text systems use notations such as ABC (Yuan et al., 2024; Wang et al., 2025). These formats differ in retained timing, grammar, sequence length, and ease of fitting a pretrained language model. The scientific question is what happens when they enter the *same* training and evaluation pipeline.

We compare seven serializers, including PMT, a compact, multi-track performance-event implementation. PMT is an engineering instance of an established event family, not an invention of fine-grained timing or velocity. It appends 609 symbols to Qwen3.5 and writes a single onset-ordered stream of time shifts, track/program changes, pitch, duration, and per-note velocity (Figure 1). PerTok supplies an independent performance-event comparator; Beat-TSD is the closest beat-grid timing comparator. The wider comparison also changes token cost and musical context per fixed block.

Auditing our earlier comparison uncovered unequal caption caps (PMT 80 text IDs, other arms up to 640), a folk-only FMD reference mislabeled as source-stratified, and train/test folk settings sharing tune IDs. We retrain all seven arms with identical 80-ID caption prefixes and evaluate on a new 500-piece reference with train-unseen folk tune IDs. Archived scores remain appendix diagnostics rather than evidence for the corrected comparison.

In our audited seven-arm grid, the raw-export margin changes sharply under message-preserving and exact common-unit protocols; the canonical note-only diagnostic makes the tested exact rescaling invariant (Section 4.3). Our evaluation audit supplies a sharper control: doubling the MIDI tick unit without changing any note time gives FMD 35.4 between playback-equivalent copies of the same 500 real scores. We therefore treat message-preserving 220-TPQ FMD as a specified embedding comparison and report raw export and other unit choices to expose metric sensitivity. In a separate 26M from-scratch study, message-preserving tick-unit controls collapse the raw PMT margin; a



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

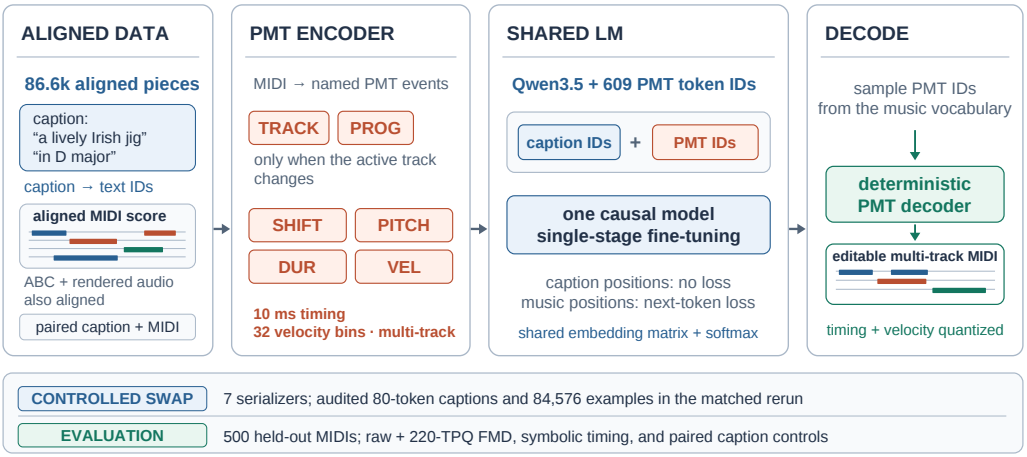


Figure 1: PMT extends the text LM with music events, masks caption loss and decodes editable MIDI. The seven-arm rerun shares 84,576 training rows and 80-token caption prefixes. Evaluation uses 100 tune-disjoint prompts and a fixed 500-piece reference; Table 1 reports raw and message-preserving 220-TPQ CLaMP-2 FMD, validity, duration and chord-time.

lossless common-unit control leaves the PMT–REMI difference unresolved. The 6.25M-pair model and published-system comparisons use different data and cannot isolate tokenizer effects. No recoverable blinded listening result supports perceptual claims.

Our contributions are an exact MIDI time-unit invariance test for CLaMP-2 FMD and an auditable multi-protocol evaluation harness; a matched seven-serializer comparison that retracts the earlier raw-export ranking; the PMT serializer as a compact implementation with a precise in-memory round-trip domain; and captioned MIDI resources with aligned ABC and rendered audio views for 86,598 pieces. We document reference MIDIs, sample counts, training IDs and checkpoint lineage behind each claim.

## 2 Related Work

**Symbolic representations.** PerformanceRNN and Music Transformer established expressive event streams with time-shift and velocity events (Oore et al., 2018; Huang et al., 2019); Anticipatory Music Transformer develops another event ordering (Thickstun et al., 2023). PMT inherits that lineage. It adds explicit per-note duration/velocity fields and track/program tags in a compact vocabulary extension for a pretrained text LLM; these are implementation choices rather than new event types. REMI aligns notes to bar positions (Huang & Yang, 2020), and MidiTok standardizes TSD, MIDI-Like and other serializers (Fradet et al., 2021). Timing and duration tokenization have been studied directly (Fradet et al., 2023b). Our MIDI-Like arm emits NoteOn, NoteOff, Velocity, TimeShift and Program tokens, making it the nearest implemented control to the PerformanceRNN event grammar. Its configured shifts use a tempo-relative lattice, whereas Oore et al. use performance-timed events. PerTok supplies an independent fine-timing control and can encode multiple instruments (Lenz & Mani, 2024). Grouped, multitrack and track-aware streams change output heads or track locality (Hsiao et al., 2021; Dong et al., 2023; Ou et al., 2025); learned music-event BPE can improve both compression and generation (Fradet et al., 2023a). Thus a result for plain ABC or one fixed-block grammar cannot establish a general rule about compact tokenization.

**Text-conditioned systems.** text2midi combines a text encoder with a REMI decoder (Bhandari et al., 2025); MIDI-LLM extends a pretrained LLM vocabulary with MIDI events (Wu et al., 2025b); MuseCoco predicts intermediate musical attributes (Lu et al., 2023); ChatMusician and NotaGen adapt text-like score formats (Yuan et al., 2024; Wang et al., 2025). Captioned MIDI resources include MidiCaps and MetaScore (Melechovsky et al., 2024; Xu et al., 2025b); automatic captioning has also been studied for audio (Doh et al., 2023). Published systems differ in data, backbone, loss and

decoding, so their head-to-head rows contextualize usefulness but cannot isolate serialization. PMT is closest architecturally to MIDI-LLM. Its empirical question differs: we rebuild several serializations on the same training IDs, text-token cap and 0.8B recipe. Audio generators such as MusicGen and MusicLM (Copet et al., 2023; Agostinelli et al., 2023) produce a different deliverable; we do not rank rendered symbolic scores against native waveforms on one audio metric.

**Evaluation.** MidiCaps supplies a captioned MIDI test domain (Melechovsky et al., 2024). FMD applies a Fréchet distance to CLaMP-2 MIDI embeddings (Retkowski et al., 2024; Wu et al., 2025a). Its original study documents sensitivity to input processing and finite-sample bias (Retkowski et al., 2024); Clean-FID likewise showed that preprocessing affects feature-distance comparisons in images (Parmar et al., 2022). Pereval tests contextual MIDI embeddings against listener ratings for expressive piano and studies a kernel distance (Gavrilev et al., 2026); that validation does not establish perceptual meaning for our CLaMP-2 FMD comparisons. Our distinct test preserves every note time, tempo map and non-time message field while rescaling tick units, then measures the effect on a controlled seven-tokenizer ranking. FMD has no paired-caption term, so we also report decoded validity, duration, symbolic marginals (Yang & Lerch, 2020) and direct caption-attribute checks. The reference mixture and export protocol are specified throughout.

# 3 Representations and Study Design

## 3.1 Question and Scope

**Problem definition.** Given a natural-language description $c$, generate a symbolic score $y$, notes with pitch, onset, duration, velocity and instrument, that is well-formed (decodable and renderable), matches the description, and matches the statistical profile of real music. A representation $R$ serializes $y$ into a token sequence with decoder $R^{-1}$; a text-to-music method is a choice of $R$ plus a procedure for training $p_\theta(R(y) \mid c)$. PMT's pipeline (Figure 1) has three interlocking components: the *representation* encodes each piece into a flat music-token stream, the *integration recipe* appends that stream to the caption and fine-tunes a pretrained backbone with the caption masked, and the *evaluation protocol* validates every design choice by training competing representations under the identical recipe. The recipe fixes caption IDs, examples, optimizer updates, and sampling. The resulting swap identifies the effect of the complete serializer under this recipe, including its music-token coverage; it does not isolate timing alone.

## 3.2 Performance-Timed Serializer

PMT is a compact multi-track instance of the performance-event lineage of Oore et al. (2018) and Huang et al. (2019), not a new kind of musical event. Its implementation uses a flat 609-symbol vocabulary: `PITCH` (128), `DURP` (200 bins of 10 ms, clipped to 10–2000 ms), `VEL` (32), `TSHIFT` (100), `TRACK` (16), `PROG` (129, including drums), and four special tokens. The useful engineering choices are explicit instrument-change markers, a single softmax for a text LLM, and a measured trade-off between timing resolution and sequence length. A flat stream exposes cross-track context but can fragment instrument-specific patterns; track-aware grouping is a reasonable alternative (Ou et al., 2025).

$$\begin{aligned}\mathcal{V}_{\text{music}} = {} & \texttt{TRACK}_{[0,16)} \sqcup \texttt{PROG}_{[0,129)} \sqcup \texttt{PITCH}_{[0,128)} \\ & \sqcup\ \texttt{DURP}_{[0,200)} \sqcup \texttt{VEL}_{[0,32)} \sqcup \texttt{TSHIFT}_{[0,100)} \sqcup \mathcal{S},\end{aligned} \tag{1}$$

where $|\mathcal{S}| = 4$ and the six disjoint parameter families plus specials total 609. For notes $(p_j, o_j, d_j, v_j, c_j, g_j)$ sorted by onset, track and pitch, $g_j$ is the instrument program and $o_0 = 0$. The encoder first rounds each *absolute* onset, $u_j = \lfloor o_j/\tau \rceil$ with $\tau = 10$ ms, then sets $\delta_j = u_j - u_{j-1}$. It emits an empty shift sequence for $\delta_j = 0$; otherwise it tiles $\delta_j$ with positive integers $s \le 100$, writing each as $\texttt{TSHIFT}_{s-1}$. Thus $\texttt{TSHIFT}_0$ means a 10 ms shift, not zero. A $\texttt{TRACK}_{c_j}$–$\texttt{PROG}_{g_j}$ pair is emitted when the active track changes, followed by $\texttt{PITCH}_{p_j}$, $\texttt{DURP}_{q_d(d_j)-1}$ and $\texttt{VEL}_{q_v(v_j)}$, where

$$\begin{aligned} q_d(d) &= \operatorname{clip}\big(\lfloor d/\tau \rceil, 1, 200\big), \\ q_v(v) &= \left\lfloor (v-1)\frac{31}{126} \right\rceil, \qquad 1 \le v \le 127. \end{aligned} \tag{2}$$

$$R_{\mathrm{PMT}}(y) = \big[\mathcal{E}(\delta_j),\ (\texttt{TRACK}_{c_j}, \texttt{PROG}_{g_j})_{\text{if changed}},\ \texttt{PITCH}_{p_j}, \texttt{DURP}_{q_d(d_j)-1}, \texttt{VEL}_{q_v(v_j)}\big]_{j=1}^{N}, \quad (3)$$

where $\mathcal{E}$ is the positive-shift tiling above. The stream has about four tokens per note. Absolute-onset rounding prevents error accumulating across notes. For scores with at most 16 instrument entries, note pitch, non-drum program and drum flag round-trip exactly and nonempty instrument-track grouping is preserved up to track renumbering; onset error is at most 5 ms, and velocity error at most two MIDI units. Duration error is at most 5 ms only for durations that do not hit the 10–2000 ms clip; longer notes can incur larger error. Appendix H states the restricted guarantee and the encode–decode steps. These bounds apply to the in-memory encode–decode score before MIDI-file export. Writing at 220 TPQ and 120 BPM additionally rounds each event to a 2.27 ms tick; the saved-file measurements in Table 1 are evaluated separately. The grid serializers' actual round trips use their own configured timing and grammar.

The comparison probes what this resolution buys. The six dedicated music vocabularies range from 444 to 609 symbols (Appendix Table 3); the same block length therefore does not fix note coverage. Beat grids quantize onset to a tempo-dependent step for tempo $B$ and $R$ steps per beat,

$$\rho = 60/(BR)\ \text{s}. \quad (4)$$

$R = 4$–$8$ gives roughly 63–125 ms at 120 BPM. PMT uses a fixed 10 ms lattice, retaining timing deviations such as swing and rubato more closely. Beat-TSD is the nearest timing control; the broader seven-way comparison also changes grammar, context coverage and track support, so its family ranking should not be attributed to timing alone. All compared MIDI tokenizers use 32 velocity levels, removing velocity-bin count as one difference between PMT and Beat-TSD. Model-free round trips quantify information loss; generation metrics then test the trained models. We treat chord-time as a target-matching statistic rather than an "expressive ceiling": a round trip above the real reference's chord-time has *overshot* the target, not surpassed it.

### 3.3 Shared LLM Training Recipe

The recipe converts a pretrained text LLM into a text-to-music generator with minimal surgery. (i) *Vocabulary extension*: the 609 music symbols are appended to the backbone tokenizer (Qwen3.5 (Qwen Team, 2026), 248k text tokens) as mean-initialized embeddings, no architecture change, one softmax. (ii) *Caption-masked SFT*: each example concatenates caption tokens $c{=}(c_1, \ldots, c_m)$ and the music stream $x{=}R_{\mathrm{PMT}}(y){=}(x_1, \ldots, x_T)$ into one 1024-token block $s$. With a per-position mask $\ell_i{=}\mathbb{1}[i{>}m]$, the backbone minimizes the *caption-masked* autoregressive negative log-likelihood over caption–score pairs $\mathcal{D}$,

$$\mathcal{L}(\theta) = -\,\mathbb{E}_{(c,y)\sim\mathcal{D}}\ \frac{1}{|\{i : \ell_i{=}1\}|}\ \sum_{i=1}^{m+T} \ell_i\ \log p_\theta\big(s_i \mid s_{<i}\big), \quad (5)$$

so gradients flow only from the $T$ music positions. Causal attention makes the caption available to music tokens, but does not prove that generated music uses it. The mask is deterministic: 100% of caption positions and 0% of music positions have their next-token loss removed; we do not claim an optimum masking proportion from a sweep. (iii) *Music-constrained decoding*: at inference, logits are renormalized over the music range only,

$$\tilde{p}_\theta(v \mid c, x_{<t})\ \propto\ p_\theta(v \mid c, x_{<t})\ \mathbb{1}[v \in \mathcal{V}_{\text{music}}]\,, \quad (6)$$

sampling until `EOS`, after which the inverse maps valid token patterns to MIDI. We separately measure whether the result contains at least eight notes; constrained vocabulary sampling alone does not ensure a useful score. The recipe has no auxiliary loss or reinforcement learning. Its tested 0.8B swap still changes music-token coverage and cannot by itself isolate the musical information retained by each format.

### 3.4 Evaluation Protocol and Model-Free Checks

We compare seven serializers in four families: performance-resolution (PMT and PerTok (Lenz & Mani, 2024)), beat-grid (Beat-TSD, REMI and MIDI-Like), a grouped Structured control, and ABC text. The split, captions, backbone, optimizer budget and sampling settings are fixed within each comparison; each serializer has its own inverse decoder. Token counts per note differ, so a

1024-token block exposes different amounts of music; only the nearest Beat-TSD control approaches a single-factor timing intervention. The conclusion is a representation effect under this protocol, not a general ranking of all compact music encodings.

For each serializer $R$, we also encode and decode the same held-out real MIDI without a model. For a target-matching distance $d$, we report the model-free round-trip distortion and the generated distance separately:

$$D_R^{\mathrm{RT}} = d\big(R^{-1}(R(Y_{\mathrm{test}})), Y_{\mathrm{test}}\big)\,, \qquad D_R^{\mathrm{gen}} = d\Big(\hat{Y}, Y_{\mathrm{test}}\Big)\,. \tag{7}$$

For chord-time, $d$ is absolute percentage-point error; for FMD, it is the CLaMP-2 Fréchet distance under the stated reference protocol. The round-trip value measures information discarded by the tokenizer, but it does not partition the generated error into "expressiveness" and "learnability." We do not use generated/round-trip chord-time ratios as quality scores or mathematical ceilings. We report inter-onset-interval and other symbolic-marginal divergences separately, and compare token losses only within one vocabulary. Archived CLaMP-3 checks on the unequal-budget grid are exploratory. FMD is a distributional proxy whose agreement with human judgment on these outputs is unvalidated; our claims therefore do not equate lower FMD with better listening quality.

## 4 EXPERIMENTS

**Data and provenance.** The aligned corpus contains 86,598 piece-level records with captions, MIDI, ABC, and FluidSynth-rendered audio views. Native ABC can describe a whole folk tune while the MIDI training target is a short slice; the matched ABC arm therefore converts exactly that MIDI slice to ABC. Four source families occupy eight source directories: folk, classical piano, a genre-balanced Lakh subset (Raffel, 2016), and jazz. Frozen splits contain 84,576 train and 2,022 test records. We normalize source MIDI, freeze the split, render score slices, caption the render with Qwen3-Omni (Xu et al., 2025a) and score metadata, then filter for language, repetition, musical content and refusals. These automatic gates do not verify caption truth. A larger caption manifest contains 6,248,214 filtered caption–MIDI rows after 6,250,143 prefilter pairs; 6,244,793 unique hash IDs remain, so we do not call it fully content-deduplicated. Its checkpoint experiment is separate from the seven-way comparison (Appendix I).

**Audited training.** Every 0.8B serializer receives the same ordered 84,576 train IDs and the identical first 80 Qwen3.5 caption-token IDs. The caption mask begins music prediction at position 80, except ABC's four-token separator puts its first music label at 84. The seven caches were audited row by row before training. All runs use 10,000 steps, effective batch 16, learning rate $8 \times 10^{-5}$ with cosine decay and 3% warmup, bf16, and 1,024-token blocks. The fraction of blocks filled with music and the number of musical seconds they cover still differ by serializer; we report this as a mechanism and limitation, not a held-fixed variable. The older 0.8B–27B grid had unequal caption budgets and is excluded from the controlled result. On the clean 100-prompt sample, every source caption exceeds 80 tokenizer IDs (mean 441); only 3 of 12 captions that literally name their metadata key retain that string in the first 80 IDs. A 320-token sensitivity condition retains 10/12, although 88/100 captions remain truncated. Thus the matched study holds a severely shortened text input fixed; its caption-control result must be interpreted for that input, not full captions.

**Generation and FMD.** Each row uses the same 100 caption indices sampled from the tune-disjoint, source-stratified 500-piece test subset with seed 0. Prompts are capped at 80 tokens. Sampling uses temperature 0.95, top-$k$ 60, a 900-new-token limit, and the representation's music vocabulary. The primary real reference is the *same fixed* 500 held-out MIDI files in `eval/clean_strat_ref500`, copied exactly from that test subset. Folk tune IDs are unique and absent from training; all 500 quantized-note fingerprints are unique and absent from the full 84,576-row training manifest (Appendix B). We also score against the old 500-piece source-stratified reference and a separate 500-piece folk-only reference as sensitivity analyses. CLaMP-2 embeds only outputs with at least eight notes, and a shared Gaussian estimator computes FMD on $N_{\mathrm{FMD}}$ valid generations versus 500 real features. We report $N_{\mathrm{FMD}}$ and the validity rate, since conditioning on success can favor a low-validity serializer; the reference choice changes absolute scale. We report duration and chord-time rather than treating FMD as a listening metric. One run per serializer does not identify seed variance.

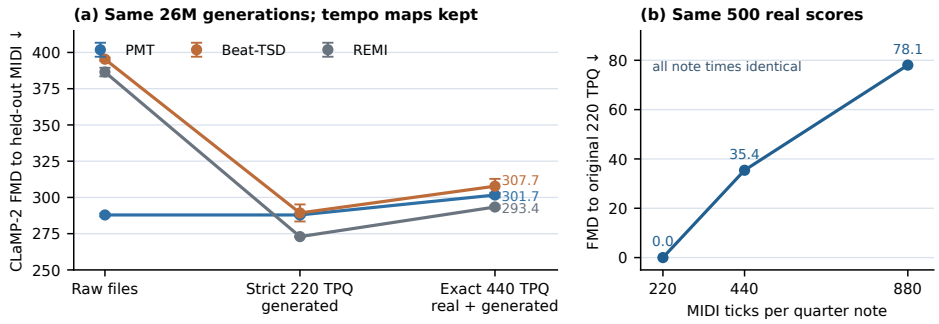


Figure 2: (a) Three-seed 26M CLaMP-2 FMD: raw exports, generated-file rescale to 220 TPQ against 202 real scores, and exact rescale of both sides to 440 TPQ. Tracks, messages and tempo maps remain; the last condition preserves every note time. Points show mean ± population SD over 198–200 files per seed. (b) The same 500 real scores rescaled losslessly to 440/880 TPQ still gain FMD. These are evaluation sensitivities, not listener scores.

### 4.1 MIDI Time-Unit Invariance Control

We first test whether the evaluation embedding changes when the music does not. For the same 500 real MIDIs, multiplying both ticks per quarter and every delta tick by two or four preserves every note onset and offset in seconds, the tempo map, and all non-time messages. Direct CLaMP-2 extraction nonetheless gives FMD 35.4 and 78.1 to the originals. These are measurement changes for identical note events. A related CLaMP-3 MIDI-as-MTF extractor gives Fréchet distance 30.6 for the exact twofold pair; this is a second embedding check, not an independent listener measure. The MTF string changes with the tick-valued fields; CLaMP-2 then groups messages into 64-character patches and 512-patch segments. The rescale changes patch counts for 144/500 and 236/500 files at twofold/fourfold units (Appendix G). Thus the observed effect belongs to the entire MTF-to-embedding path, not the TPQ header alone. They motivate a declared message-preserving 220-TPQ primary protocol and separate raw and lossless common-unit sensitivity scores; normalization cannot validate the embedding against listeners.

Source-specific nulls appear in Appendix Table 8. **From-scratch diagnostic.** A separate three-seed 26M control gives raw PMT/Beat-TSD/REMI FMD 287.9/395.3/386.5 against 202 held-out scores; exports use 220/8/8 TPQ. Message-preserving rescaling to 220 TPQ gives PMT/Beat-TSD/REMI FMD 287.9/289.3/273.0 (maximum note-time shift 2.92 ms). Lossless common 440-TPQ rescaling gives 301.7/307.7/293.4. The exact-unit PMT-minus-REMI difference is 8.3 (bootstrap 95% interval [-1.8,14.3]). Both common-unit controls collapse the raw PMT gap; the exact PMT–REMI interval spans zero. Figure 2 and Appendix F give the full protocol and per-seed scores.

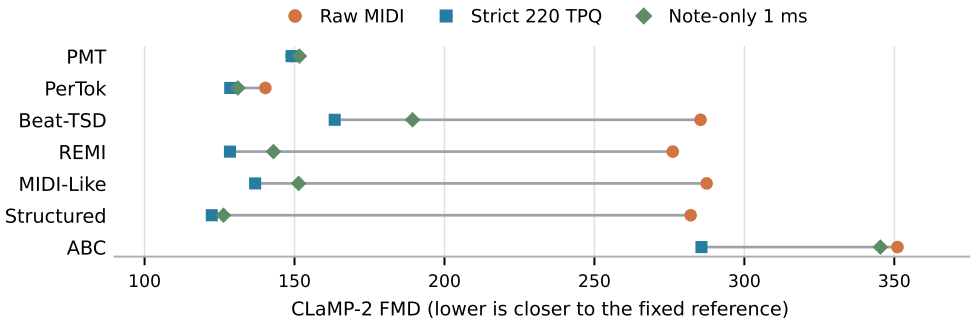


Figure 3: The same seven seed-0 generation sets under raw MIDI, message-preserving 220-TPQ, and canonical note-only 1-ms CLaMP-2 protocols. The corresponding real reference always has 500 scores. Lines connect raw and strict scores; diamonds show the separate note-only projection. These changes measure protocol sensitivity, not listener preference.

### 4.2 MATCHED REPRESENTATION COMPARISON

| Serializer | 220-TPQ FMD↓ | Raw FMD↓ | IF1↑ | $N_{\mathrm{FMD}}$ | Valid | Sec. | Chord% |
|---|---|---|---|---|---|---|---|
| PMT | 149.0 | 149.0 | 0.730 | 98 | 0.98 | 32.3 | 30.3 |
| PerTok | 128.6 | **140.3** | 0.808 | **100** | **1.00** | 27.2 | 28.4 |
| Beat-TSD | 163.4 | 285.4 | 0.741 | 98 | 0.98 | 30.9 | **32.7** |
| REMI | 128.5 | 276.1 | 0.801 | **100** | **1.00** | 32.6 | 31.1 |
| MIDI-Like | 136.9 | 287.4 | 0.804 | 99 | 0.99 | 29.3 | 28.3 |
| Structured | **122.4** | 282.1 | **0.817** | **100** | **1.00** | 32.1 | 29.5 |
| ABC | 285.7 | 351.1 | 0.776 | 98 | 0.98 | **36.9** | 20.1 |
| Real prompts (100) | 68.9 | 68.9 | 0.809 | 100 | – | 41.3 | 33.7 |
| Real reference | 0 | 0 | – | 500 | – | 42.1 | 35.0 |

Table 1: Seven seed-0 runs on 84,576 shared slices, 80 caption IDs, 10,000 updates and a 900-token decoder cap. CLaMP-2 FMD uses one clean 500-piece reference; the strict column rescales generated files to 220 TPQ while retaining messages and tempo maps. IF1 is coarse GM-family F1 against the visible caption, with invalid outputs scored as empty. $N_{\mathrm{FMD}}$ counts outputs with at least eight notes; Valid is out of 100. Sec. and chord-time use saved valid files. The 100 real prompts belong to the reference and calibrate finite-sample FMD. None is a listener score.

Under the specified message-preserving 220-TPQ protocol, the saved PMT/Beat-TSD outputs score FMD 149.0/163.4, compared with raw 149.0/285.4 (Table 1). The Beat-TSD-minus-PMT margin falls from 136.4 to 14.4; the transform keeps all tempo messages and categorical note events while shifting note times by at most 1.14 ms. Structured (122.4), REMI (128.5) and PerTok (128.6) have lower point FMD than PMT (149.0) on this single-seed sample; their verified six-comparison simultaneous 95% intervals are entirely below zero; Beat-TSD's interval spans zero. Folk-only raw FMD is a separate reference sensitivity. These are embedding-distribution comparisons under the stated export protocol, not listener judgments or a timing-only intervention.

**Matched-output uncertainty.** A 1,000-replicate bootstrap resamples shared prompt indices and the common real set; max-standardized 95% intervals cover the six PMT contrasts simultaneously. Competitor-minus-PMT strict FMD and interval are Beat-TSD 14.4 [-1.6,30.4]; Structured -26.6 [-45.1,-8.1]; REMI -20.5 [-39.7,-1.4]; PerTok -20.4 [-40.3,-0.6]. These are single-training-seed sampling intervals, not listener or model-training uncertainty.

**Finite-sample control.** Even a 100-piece subsample of the real 500 has nonzero FMD to that same reference: 68.9 for the frozen prompt indices; 1,000 random real subsamples give mean 68.1 (2.5–97.5th percentiles 65.0–72.3). The subset is included in the reference, so this is a sampling calibration, not a universal lower bound or a listener baseline.

**Interpretation.** Beat-TSD is the nearest timing control: like PMT it uses 32 velocity levels, but its tempo-relative time grid and token grammar differ. PerTok independently encodes fine timing, per-note velocity and multiple instruments; its configured serializer round-trips a seven-instrument source example without losing instrument count. Thus a PMT–PerTok difference must not be described as a track-support gain. An archived round-trip comparison used a separate PMT grid-mode serializer on a folk-only 200-piece reference; those values are not Beat-TSD round trips and cannot apportion this generative gap into expressiveness and learnability. Fixed token blocks also expose different music lengths. The matched result is a serializer-family observation under a single optimization budget.

**Model-free check.** All seven serializers decode 500/500 real scores. Raw round-trip FMD is 12.0 for PMT and 271.4 for Beat-TSD; re-exporting decoded notes at 220 TPQ/120 BPM gives 12.1/41.8. This re-export also rewrites MIDI events and can flatten tempo changes, so neither difference partitions representability from learnability. A narrower message-preserving Beat-TSD 8-to-220-TPQ rescale gives FMD 173.1 with at most 1.12 ms note-time shift. Full results are in Appendix D.

**Independent symbolic advantages.** On raw decoded notes, PMT has lower inter-onset, duration and velocity JSD than Beat-TSD: 0.045/0.070/0.016 versus 0.189/0.139/0.023. PMT is lowest among the seven arms on duration and velocity JSD, while PerTok is lowest on inter-onset JSD (0.008). These distributional marginals do not measure coherence or listener preference (Appendix D).

**Lossless common-unit stress test.** Scaling both real and generated tick units by integer factors to 10,560 TPQ preserves every note time, tempo message and non-time event. PMT/Beat-TSD/REMI/PerTok/Structured FMD becomes 107.4/128.5/106.8/98.6/98.8; all seven values are in Appendix G. The PMT–REMI point gap nearly vanishes. This unusually fine unit tests protocol stability, not a preferred metric setting. The six-comparison simultaneous 95% intervals are Beat-TSD-minus-PMT [7.2,34.9] and REMI-minus-PMT [-21.8,20.5].

### 4.3 Canonical Note-Only FMD

We keep instrument order, program, drum flag, pitch, velocity and note times, round absolute onsets/offsets to 1 ms, and write fixed 500-TPQ/120-BPM MIDI. Tempo maps, controllers, signatures and other non-note events are dropped. All 500 real and 693 generated files become byte-identical to their exact integer tick-rescaled counterparts. Thus the same-music null is zero after this note-only projection. Table 2 reports every arm and six fixed-checkpoint contrasts. This tested invariance does not imply equivalence under arbitrary MIDI edits or listener validity.

| Serializer | FMD ↓ | Minus PMT, simultaneous 95% CI |
|---|---|---|
| PMT | 151.6 | reference arm |
| PerTok | 131.1 | -20.5 [-39.5,-1.5] |
| Beat-TSD | 189.4 | +37.7 [19.0,56.5] |
| REMI | 143.0 | -8.6 [-29.9,12.6] |
| MIDI-Like | 151.4 | -0.3 [-22.1,21.6] |
| Structured | **126.3** | -25.3 [-45.1,-5.5] |
| ABC | 345.4 | +193.7 [156.3,231.2] |

Table 2: Fresh CLaMP-2 FMD on canonical note-only MIDI for the same 500 real scores and 98–100 fixed seed-0 outputs per serializer. Six PMT contrasts use one shared-prompt/reference bootstrap with max-standardized simultaneous 95% sampling intervals; they omit training variability.

**Targeted tempo-token control.** We trained an extra Beat-TSD arm with tempo tokens disabled on the identical 84,576 rows, 80-token caption prefixes, 10,000 updates and decoding budget. All valid decoded outputs carry a 120-BPM header; the original Beat-TSD uses 121.29 BPM. The extra arm scores strict 220-TPQ clean-reference FMD 113.3 (raw 270.9) and folk-reference raw FMD 319.8 with 100/100 valid outputs and mean valid duration 33.8 seconds. Its independent IOI JSD is 0.142. Removing the tempo token also changes the music vocabulary and block coverage; this is a targeted configuration sensitivity, not a single-parameter timing proof.

**PMT versus the nearest beat-grid control.** In the canonical note-only comparison, PMT scores 151.6 versus Beat-TSD's 189.4 FMD: Beat-minus-PMT is 37.7 with a simultaneous 95% sampling interval [19.0,56.5] for fixed seed-0 checkpoints. PMT also has lower standardized model-free round-trip FMD (12.1 versus 41.8). Its 1,024-token block truncates 33.4% of music streams versus 39.3% for Beat-TSD; serializer grammar and context coverage remain entangled (Appendix Table 3).

**Reference distribution.** The fixed 500-piece set contains 300 folk, 132 balanced, 62 classical and 6 jazz scores (mean 42.1 seconds, 2.03 tracks, 35.0% chord-time); a separate 500-piece folk reference has only 9.5% chord-time. FMD is the Fréchet distance between Gaussian means and covariances of CLaMP-2 features. The same real array is used for all serializers within a column, and the 100 caption indices (57/33/9/1 by source) are stored in every manifest. FMD has no paired-caption term.

**Split repair.** An earlier setting-level folk split left 895/1,216 test settings with train-seen tune IDs, and our first proposed reference had 19/500 exact train fingerprints. Before the matched rerun, we rebuilt the fixed 500-piece set with unseen folk tune IDs, unique source paths and zero exact train fingerprints. This does not exclude transpositions or near-duplicate arrangements (Appendix B).

Across the seven matched arms, Table 1 also reports visible-caption coarse instrument-family F1. PMT/Beat-TSD/PerTok score 0.730/0.741/0.808; the paired PMT-minus-Beat-TSD difference is -0.011 (95% bootstrap interval [-0.053, 0.028]). PMT-minus-PerTok is -0.078 ([-0.125,-0.036]); PerTok is higher on this caption-attribute metric. The corresponding 100 real MIDIs score 0.809 against their machine captions, indicating that this lexical/program proxy has label mismatch even on

the source scores. On the 33 multi-instrument balanced prompts, real/PMT/Beat-TSD/PerTok F1 is 0.523/0.343/0.368/0.531. It tests the visible 80-token prefix, not full-caption or harmonic adherence.

### 4.4 Caption Conditioning Control

We keep checkpoint, prompt indices, decoder and per-prompt seeds fixed while changing only the visible 80-token caption prefix. Coarse GM-family F1 uses the original visible prefix as target; invalid MIDI scores zero. Original/cross-source shuffled/neutral F1 is 0.730/0.551/0.346; the original-minus-cross-source shuffle interval is [0.103,0.254]. But swapping similar-length captions within source changes F1 by only 0.015 over 99 prompts (95% paired interval [-0.022,0.054]); among 49 pairs with different visible family sets, the interval is [-0.043,0.081]. Thus the cross-source result does not establish specific instrument control. A 11-prompt one-word instrument edit changes requested-family presence from 0.182 to 0.545 (difference interval [0.091,0.636]). The selected edit probes local lexical response, not reliable caption control. Neither probe establishes key, chord-sequence or perceptual alignment (Appendix C).

**Longer-caption sensitivity.** On the same source rows and 10,000-update schedule, raising the caption cap from 80 to 320 IDs yields PMT/Beat-TSD strict 220-TPQ FMD 117.2/130.3 (raw 117.2/273.5) with 100/100 valid outputs. The PMT original/shuffled/neutral prompt control obtains visible-prefix coarse instrument-family F1 0.640/0.480/0.476; its strict 220-TPQ FMD values are 117.2/114.3/274.0. The 320-token cap retains literal requested-key text in 10/12 relevant sampled captions, but still truncates 88/100 captions. The audited cap-320 caches match on caption IDs and first music position, but 45.5% of PMT blocks and 75.5% of Beat-TSD blocks end with a music label. This is a sensitivity test rather than a controlled isolation of caption length.

**Reference and decode sensitivity.** The same 165 outputs flip PMT-6M/Amadeus FMD from 313.5/223.9 on full-score references to 177.4/297.7 on 256-note prefixes. This changes real length and tempo, and 159/165 PMT targets occur in training: it diagnoses reference sensitivity, not generalization (Appendix J). Raising a PMT-4B cap from 900 to 7,000 tokens increases median duration from 18.25 to 44.27 seconds and changes FMD (Appendix L).

**Claims by protocol.**

| Check | Held fixed | Supported inference |
|---|---|---|
| Exact rescale | Notes, seconds, messages | FMD unit sensitivity |
| Strict 220 TPQ | Messages, tempo; $\leq$1.14 ms drift | Common-unit FMD ranking |
| Note-only 1 ms | Rounded note seconds, pitch, velocity, program | Projected FMD ranking |

Note-only projection drops non-note events. FMD needs a declared reference, export and valid count; duration, captions and listeners need separate checks (Figure 3).

### 4.5 Limitations and Perceptual Scope

Primary arms have one 10,000-update seed, unequal music coverage per block and unverified convergence. One hundred outputs are small for 768-dimensional FMD; paired intervals omit retraining variability and changes in reference composition. CLaMP-2 FMD is export-sensitive and unvalidated against listening here; no recoverable blinded listening result exists. The reference has six jazz pieces. Captions lack a factual audit, 80-ID prefixes omit content, and key, chords and requested duration remain weak or untested. Deduplication misses near relatives; the archived scale grid used unequal caption caps.

## 5 Conclusion

In the 0.8B grid, PMT's raw FMD lead over Beat-TSD (136.4) shrinks to 14.4 under 220-TPQ export (interval crosses zero), while the note-only gap is 37.7 [19.0,56.5]. Structured and PerTok score better than PMT canonically. Exact same-music rescaling gives FMD 35.4, exposing metric sensitivity. One seed, unequal context and weak caption/listener validation restrict claims; 6.25M is a training-exposed scale probe.

### AI USE STATEMENT

Generative AI was used for grammar correction and audio-grounded training captions, and assisted code, analysis, manuscript drafting, and figures. The authors verified reported results against stored artifacts and take responsibility for the final paper; Appendix I documents captioning.

### ETHICS STATEMENT

Source music was obtained under source-specific terms, including purchased AI-training rights asserted by the data custodian. Public redistribution requires a separate rights check. No unverified listener result enters our claims.

### REPRODUCIBILITY STATEMENT

The anonymous supplement contains a listening gallery, feature arrays, hashes and CPU score replay, plus 21 generated-MIDI pairs for independent tick-rescale and feature-extraction checks. Source reference MIDI are withheld pending redistribution review, and checkpoints exceed the upload limit; training and the 500-piece source intervention cannot be fully reproduced from the public package.

## REFERENCES

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. MusicLM: Generating music from text. *arXiv:2301.11325*, 2023.

Keshav Bhandari et al. text2midi: Generating symbolic music from captions. In *AAAI*, 2025.

Louis Bradshaw and Simon Colton. Aria-MIDI: A dataset of piano MIDI files for symbolic music modeling. In *International Conference on Learning Representations (ICLR)*, 2025.

Jade Copet, Alexandre Défossez, et al. Simple and controllable music generation. In *NeurIPS*, 2023.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. LP-MusicCaps: LLM-based pseudo music captioning. In *ISMIR*, 2023.

Hao-Wen Dong, Ke Chen, Shlomo Dubnov, Julian McAuley, and Taylor Berg-Kirkpatrick. Multitrack music transformer. In *ICASSP*, 2023.

Jeff Ens and Philippe Pasquier. Mmm: Exploring conditional multi-track music generation with the transformer. In *arXiv:2008.06048*, 2020.

Jeff Ens and Philippe Pasquier. Building the MetaMIDI dataset: Linking symbolic and audio musical data. In *ISMIR*, 2021.

Nathan Fradet, Jean-Pierre Briot, Fabien Chhel, Amal El Fallah Seghrouchni, and Nicolas Gutowski. Miditok: A python package for midi file tokenization. In *ISMIR Late-Breaking Demo*, 2021.

Nathan Fradet, Nicolas Gutowski, Fabien Chhel, and Jean-Pierre Briot. Byte pair encoding for symbolic music. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2001–2020, 2023a. doi: 10.18653/v1/2023.emnlp-main.123.

Nathan Fradet, Nicolas Gutowski, Fabien Chhel, and Jean-Pierre Briot. Impact of time and note duration tokenizations on deep learning symbolic music modeling. In *ISMIR*, 2023b.

Dmitrii Gavrilev et al. Integrating contextual embeddings into evaluation of expressive MIDI piano performances. In *Proceedings of ISMIR*, 2026.

Curtis Hawthorne et al. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *Proc. ICLR*, 2019.

Wen-Yi Hsiao, Jen-Yu Liu, Yin-Cheng Yeh, and Yi-Hsuan Yang. Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 178–186, 2021.

Cheng-Zhi Anna Huang et al. Music transformer: Generating music with long-term structure. In *ICLR*, 2019.

Yu-Siang Huang and Yi-Hsuan Yang. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *ACM Multimedia*, 2020.

Qiuqiang Kong, Bochen Li, Jitong Chen, and Yuxuan Wang. GiantMIDI-Piano: A large-scale MIDI dataset for classical piano music. *arXiv:2010.07061*, 2020.

Keon Ju Maverick Lee et al. The GigaMIDI dataset with features for expressive music performance detection. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, 2025.

Julian Lenz and Anirudh Mani. Pertok: Expressive encoding and modeling of symbolic musical ideas and variations. In *Proc. ISMIR*, 2024.

Shuyu Li, Dooho Choi, and Yunsick Sung. MIDILM: A dual-path model for controllable text-to-MIDI generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(28):23160–23168, 2026.

Peiling Lu et al. Musecoco: Generating symbolic music from text. arXiv:2306.00110, 2023.

Jan Melechovsky, Abhinaba Roy, and Dorien Herremans. Midicaps: A large-scale midi dataset with text captions. In *ISMIR*, 2024.

Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan. This time with feeling: Learning expressive musical performance. In *Neural Computing and Applications*, 2018.

Longshen Ou, Jingwei Zhao, Ziyu Wang, Gus Xia, Qihao Liang, Torin Hopkins, and Ye Wang. Unifying symbolic music arrangement: Track-aware reconstruction and structured tokenization. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *Proceedings of CVPR*, 2022.

Xingwei Qu et al. Mupt: A generative symbolic music pretrained transformer. In *ICLR*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents. `https://huggingface.co/Qwen/Qwen3.5-0.8B`, 2026.

Colin Raffel. Learning-based methods for comparing sequences, with applications to audio-to-MIDI alignment and matching. *PhD thesis, Columbia University*, 2016.

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

Lingyu Su et al. Amadeus: Autoregressive model with bidirectional attribute modelling for symbolic music. *arXiv preprint arXiv:2508.20665*, 2025.

John Thickstun, David Hall, Chris Donahue, and Percy Liang. Anticipatory music transformer. arXiv:2306.08620, 2023.

Dimitri von Rütte, Luca Biggio, Yannic Kilcher, and Thomas Hofmann. Figaro: Controllable music generation using learned and expert features. In *ICLR*, 2023.

Yashan Wang et al. Notagen: Advancing musicality in symbolic music generation with large language model training paradigms. In *IJCAI*, 2025.

Ziyu Wang et al. POP909: A pop-song dataset for music arrangement generation. In *Proc. ISMIR*, 2020.

Shangda Wu et al. CLaMP 2: Multimodal music information retrieval across 101 languages using large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 435–451, 2025a. doi: 10.18653/v1/2025.findings-naacl.27.

Shih-Lun Wu et al. Midi-llm: Adapting large language models for text-to-midi music generation. In *NeurIPS Workshop on AI for Music*, 2025b.

Jin Xu et al. Qwen3-Omni technical report. *arXiv:2509.17765*, 2025a.

Weihan Xu, Julian McAuley, Taylor Berg-Kirkpatrick, Shlomo Dubnov, and Hao-Wen Dong. Generating symbolic music from natural language prompts using an LLM-enhanced dataset. In *ISMIR*, 2025b.

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

Ruibin Yuan et al. Chatmusician: Understanding and generating music intrinsically with llm. In *Findings of ACL*, 2024.

Mingliang Zeng et al. Musicbert: Symbolic music understanding with large-scale pre-training. In *Findings of ACL*, 2021.

## A Evaluation Protocol Checklist

Publish exact reference and prompt IDs, exported MIDI, validity rules, saved embeddings and row maps. Before interpreting model gaps, score a lossless same-music tick-unit null; then show a message-preserving common-unit score and an exact second unit. Audit note times, messages, patch counts and hashes. Report valid counts, durations and direct caption attributes; pair model contrasts by prompt and separate sampling from training-seed uncertainty. File-unit invariance does not establish listener preference.

## B Audited Protocol and Cache Checks

All seven cap-80 training caches contain 84,576 ordered rows from the same source manifest. Their caption-ID prefixes match exactly row by row. PMT, Beat-TSD, REMI, MIDI-Like, PerTok and Structured begin music labels at position 80; ABC uses a four-token separator and begins at 84. The audit JSON records per-arm block-filling fractions. The primary 500 real MIDIs are copies of the source-stratified caption subset; their folk tune IDs have zero overlap with training, all 500 paths differ from train paths, and every reference quantized-note fingerprint is unique and absent from 81,932 unique training fingerprints. The fingerprint is SHA1 of the sorted pitch, 10 ms onset and 10 ms duration triples; this exact-match test does not reject transpositions or arrangement-level relatives. The archived stratified reference had 218/301 folk test settings whose tune ID appeared in training. We fix the training seed to zero and report a single run per arm. Every generated file and its source caption index is retained with the evaluator manifest. An eight-note minimum defines validity; FMD excludes failures and reports its actual generated count. Full command arguments are in the matched evaluation scripts and manifests. All final FMD scores use frechet-music-distance 1.0.0 with direct CLaMP-2 feature extraction: its bundled joblib feature cache keys on directory path alone, so reusing a path after rebuilding files can silently reuse stale embeddings. The primary scorer additionally fingerprints the exact MIDI bytes before reading its own cached feature array. Symbolic JSD bins use log-spaced 0.02–4-s inter-onset and duration edges (30 each), 32 velocity bins, 12 pitch classes, 20 density bins over 0–20 notes/s, and 14 polyphony-proxy bins over 1–8. Every axis has underflow and overflow bins; IOI excludes intervals at or below 5 ms as simultaneous or near-simultaneous events. The score artifact records per-arm piece, note and eligible-interval counts. Within-piece histograms are normalized before averaging across pieces; density and polyphony give each piece one observation.

| | PMT | Beat-TSD | REMI | MIDI-Like | PerTok | Structured | ABC |
|---|---|---|---|---|---|---|---|
| Music symbols | 609 | 475 | **444** | 562 | 473 | **444** | – |
| Music label start | 80 | 80 | 80 | 80 | 80 | 80 | 84 |
| Mean music labels | 715 | 805 | 818 | 874 | 904 | 808 | **567** |
| Block reaches end (%) | **33.4** | 39.3 | 40.8 | 66.2 | 79.3 | 40.8 | 42.7 |

Table 3: Direct cache audit on the same 84,576 MIDI slices and first 80 caption IDs. ABC uses the backbone text vocabulary and a four-token separator. "Block reaches end" is the fraction truncated within music at 1,024 tokens; token cost changes musical context.

## C Within-Source Caption Control

For each source family, we sort the frozen 100 prompts by full-caption token length and swap adjacent captions; the final three in an odd-sized family rotate. This preserves source and approximately matches caption length. The one jazz prompt has no within-source partner and is excluded. The same PMT checkpoint, original per-prompt seed, visible 80-token cap and decoder are used. Invalid outputs count as empty. All F1 targets refer to the original visible prefix; the table also isolates pairs whose visible instrument-family sets actually differ. A bootstrap resamples the paired prompt indices 10,000 times. Source-specific or instrument-word effects are not equivalent to exact key, chord-sequence or perceptual following.

**One-word counterfactual.** Among 11 balanced prompts with exactly one literal instrument word in the visible prefix, we replace that one word with a different GM family while keeping the frozen

| Subset | $n$ | Original F1 | Within-source F1 | Paired difference [95% CI] |
|---|---|---|---|---|
| All 99 nonself | 99 | **0.728** | 0.712 | 0.015 [-0.022,0.054] |
| Changed family set | 49 | **0.498** | 0.483 | 0.016 [-0.043,0.081] |
| Balanced | 33 | **0.343** | 0.279 | 0.063 [-0.034,0.166] |
| Classical | 9 | 0.748 | **0.774** | -0.026 [-0.222,0.148] |
| Folk | 57 | 0.947 | **0.953** | -0.006 [-0.018,0.000] |

Table 4: Within-source caption permutation using saved PMT MIDI. The confidence intervals condition on one trained checkpoint and one set of decoder seeds.

| Serializer | TPQ | BPM | Raw FMD↓ | Std. FMD↓ | \|Δnotes\| | \|Δsec.\| |
|---|---|---|---|---|---|---|
| PMT | 220 | 120.00 | **12.0** | **12.1** | **0.0** | **0.01** |
| PerTok | 320 | 120.00 | 195.1 | 15.8 | 0.1 | **0.01** |
| Beat-TSD | 8 | 121.29 | 271.4 | 41.8 | 0.1 | 0.44 |
| REMI | 8 | 121.29 | 271.4 | 41.8 | 0.1 | 0.44 |
| MIDI-Like | 8 | 121.29 | 271.2 | 42.0 | 0.1 | 0.44 |
| Structured | 8 | 120.00 | 277.7 | 20.4 | 0.1 | 0.19 |
| ABC | 480 | 120.00 | 263.1 | 43.4 | 31.3 | 2.48 |

Table 5: Model-free encode–decode audit on the 500 tune-disjoint source MIDIs of Table 1; all seven serializers decode 500/500. TPQ is MIDI ticks per quarter note; all real files use 220 TPQ and 120 BPM. Raw FMD uses each serializer's exported MIDI; standardized FMD re-exports the same decoded notes at 220 TPQ and 120 BPM. Note identities are unchanged and time shifts are at most 1.2 ms. The last two columns compare each raw reconstruction with its source piece. ABC additionally depends on midi2abc/abc2midi compilation. Neither FMD column is a listener score or a mathematical generative ceiling.

prompt index, checkpoint, decoder and sampling seed. Presence of the newly requested family changes from 0.182 to 0.545; presence of the replaced family changes from 0.727 to 0.545. The list of eligible indices, exact before/after words and per-file outcomes is retained in the plan and analysis JSON. This is a small balanced-source lexical control, not a test of full-caption understanding.

## D Model-Free Reconstruction and Symbolic Marginals

**Model-free reconstruction.** We encode and decode each of the same 500 clean-reference MIDIs with the exact seven serializers used in the trained comparison, before fitting any model. For each successful reconstruction, the real FMD comparator uses the corresponding source indices; failures are counted separately. This tests representable symbolic information under the configured quantizer and inverse, not generator quality.

With the same 220-TPQ/120-BPM export, PMT, PerTok, Beat-TSD and ABC have model-free FMD 12.1, 15.8, 41.8 and 43.4, respectively. The much larger raw distances for PerTok and the beat-grid arms therefore mix musical reconstruction with file-export differences. The residual standardized gap is descriptive: the tokenizers also differ in quantization, grammar, and, for ABC, compilation.

Disabling Beat-TSD tempo tokens reconstructs 500/500 scores at 120 BPM but still has raw FMD 277.5 at eight TPQ: tempo-bin mismatch alone does not explain the gap.

Beat-TSD exports at 8 TPQ and 121.29 BPM versus the real 220/120. Re-exporting its decoded notes at 220 TPQ and the original initial BPM changes FMD 271.4→33.7; setting 120 BPM gives 41.8. The pinned FMD MIDI-as-MTF loader prefixes the `ticks_per_beat` field, exposing file units to the embedder.

Across 165,113 notes, the same-initial-BPM re-export preserves track, program, drum flag, pitch and velocity; note times move by at most 1.12 ms.

A stricter 8-to-220-TPQ tick rescale retains all non-time messages, tracks, order and BPM: FMD changes 271.4→173.1 with at most 1.12 ms note-time shift. The larger general re-export effect also includes rewritten event serialization.

| Serializer | IOI | Velocity | Duration | Pitch class | Density | Poly. proxy |
|---|---|---|---|---|---|---|
| PMT | 0.045 | **0.016** | **0.070** | **0.000** | 0.041 | 0.037 |
| PerTok | **0.008** | 0.046 | 0.083 | 0.004 | **0.039** | 0.026 |
| Beat-TSD | 0.189 | 0.023 | 0.139 | 0.001 | 0.064 | 0.038 |
| REMI | 0.184 | 0.037 | 0.150 | 0.002 | **0.039** | 0.037 |
| MIDI-Like | 0.174 | 0.033 | 0.163 | 0.003 | 0.044 | **0.021** |
| Structured | 0.146 | 0.047 | 0.121 | 0.001 | 0.068 | 0.028 |
| ABC | 0.075 | 0.849 | 0.127 | 0.002 | 0.149 | 0.074 |

Table 6: Independent symbolic JSD from the raw decoded MIDI in seconds to the final clean 500-piece reference (lower means closer distribution). IOI is the inter-onset interval above 5 ms; density is notes per second; the polyphony proxy is note count divided by unique rounded-onset count. Explicit underflow and overflow bins retain values outside finite bin boundaries. Histograms are normalized per piece before averaging so long scores do not dominate. FMD and marginal JSD are different measurements of distribution fit, not human judgments.

Finally, doubling the tick header and every delta tick for the *same* 500 real files preserves all 165,141 notes at identical times and every non-time message, yet gives FMD 35.4 to the originals (matched embedding cosine 0.939). This is a direct failure of MIDI time-unit invariance.

The exact fourfold rescale to 880 TPQ gives FMD 78.1 under the same audit.

CLaMP-3 embeddings on this exact 220-to-440-TPQ pair also shift (Fréchet distance 30.6, mean paired cosine 0.912); both models use a MIDI-as-MTF front end. Re-extracting the original 500 files gives identical CLaMP-3 features, ruling out extractor randomness in this control.

**Symbolic marginal checks.** To keep CLaMP-2 from carrying the whole interpretation, we independently compare per-piece normalized MIDI histograms with Jensen–Shannon divergence (JSD, bits) on six interpretable axes. Each arm uses the same valid-output set as FMD and the same 500 real files. These statistics show which marginal is matched; no single marginal certifies musical coherence or caption use.

For IOI JSD, PMT scores 0.045 (95% generated-piece bootstrap [0.035, 0.068]) and Beat-TSD 0.189 ([0.171, 0.220]); the 500 real reference pieces are held fixed in this interval. Differences across axes prevent reducing the comparison to a single quality ranking.

**MIDI-export sensitivity.** A decoded-note re-export at 220 TPQ retaining each file's initial BPM gives PMT/Beat-TSD/PerTok FMD 149.4/163.5/128.7; setting 120 BPM gives 149.4/181.7/128.7. All categorical note attributes survive both re-exports; largest recorded note-time change is 1.14 ms. The re-export flattens later tempo changes and rewrites event serialization. Both are specified counterfactual protocols, not a privileged measure of listener quality; seven-arm values are retained with the outputs.

## E Legacy Grid and Reference Sensitivity

The archived seven-way 0.8B grid reported PMT FMD 159 and Beat-TSD approximately 286. Those numbers used a 500-piece folk-only reference, not the source-stratified set described in the previous draft. Its PMT training cache had an 80-token caption cap, while the other tokenizer caches admitted up to 640. A PMT seed-0 run named `q35_08b_tsd_full` also lacks an independently verified train-only cache lineage. For these reasons we do not use that grid to support a controlled or cross-scale claim. Re-scoring old generated files on the source-stratified reference gives PMT 110.5, PerTok 141.0, Beat-TSD 270.2, REMI 266.3 and MIDI-Like 275.7; these are descriptive old-generation measurements with the same caption-budget caveat. The new Table 1 supersedes this grid. A common-grid onset snap and CLaMP-3 comparison exist only on the old outputs, so they are exploratory. On the same archived 100 Beat-TSD outputs, re-exporting at 220 ticks per quarter while retaining 121.29 BPM changes clean500 FMD from 271.3 to 123.9; subsequently setting 120 BPM gives 132.6. This is a within-output export-sensitivity check, not a corrected controlled model score.

## F FROM-SCRATCH MIDI-EXPORT DIAGNOSTIC

The separate unconditional control uses an eight-layer, 512-hidden-width, eight-head 26M decoder-only Transformer with 1,024-token blocks. It trains for 2,500 optimizer steps on 1,208 multi-instrument MIDIs and samples up to 200 scores per seed for each of PMT, Beat-TSD and REMI. The real reference has 202 MIDIs; all its quantized-note fingerprints are distinct and absent from the training files. Saved generation counts are PMT 198/199/200, Beat-TSD 198/199/198 and REMI 200/200/200 across seeds 0/1/2. The raw CLaMP-2 FMD scores were recomputed without the installed package's directory-only feature cache. The PrettyMIDI export diagnostic rewrites decoded notes at 220 TPQ with either the original *initial* BPM or 120 BPM while retaining note seconds, instruments, key/time signatures, control changes and pitch bends. It flattens any later tempo changes and may rewrite event ordering; it does not isolate TPQ alone. The PMT self re-export checks side effects for outputs already carrying 220 TPQ/120 BPM. The normalized files are counterfactuals, not newly trained samples. Scripts and values are in `pmt/score_fs_tempo120_all.py` and `eval/tempo_metadata_control/fs_all/scores.json`. The strict 220-TPQ transform preserves every message's type and content, track and order; largest note-time difference is 2.92 ms across all nine sets because some outputs contain tempo changes. The exact 440-TPQ transform has no rounding and preserves the full reference and generated tempo maps. Per-file audit and feature arrays are in `eval/tempo_metadata_control/fs_mido_220tpq` and `eval/tempo_metadata_control/fs_exact_common440`. The six independent symbolic marginal histograms are cellwise identical before and after exact 440-TPQ rescaling for the reference and all nine generation sets; see `symbolic_invariance.json` in the latter directory.

| Serializer | Raw FMD, seeds 0/1/2 | Strict 220 TPQ | Exact common 440 TPQ |
|---|---|---|---|
| PMT | **289.0/286.5/288.2** | 289.0/286.5/288.2 | 303.4/299.7/301.9 |
| Beat-TSD | 396.3/393.8/396.0 | 295.6/290.8/281.5 | 314.4/306.8/302.1 |
| REMI | 388.4/388.7/382.3 | **273.5/273.0/272.4** | **294.8/292.6/292.8** |

Table 7: The same 26M output sets. Strict 220 changes only the tick unit, rounding ticks where required, and retains the complete tempo map and every non-time message. Exact common 440 rescales both real and generated MIDI by integral factors with identical note times. All columns use direct, uncached CLaMP-2 extraction and the corresponding 202-piece real feature array.

## G LOSSLESS MIDI TICK-UNIT CONTROL

For each clean-reference MIDI we multiply the TPQ header and every track-local delta tick by two or four with mido, preserving every non-time message, track and order. Within any tempo segment, elapsed seconds are $\Delta t_{\text{tick}} T/(10^6\ \text{TPQ})$ for tempo $T$ in microseconds per beat; multiplying both tick count and TPQ by the same integer leaves time unchanged even across tempo changes. The 500 originals and copies have identical 165,141 note onset/end times in seconds. FMD to the same originals is 35.4/78.1 at 440/880 TPQ, versus zero without the file-unit change. The effect remains positive separately for folk ($n = 300$; 40.7/84.1), balanced ($n = 132$; 50.1/93.5), and classical ($n = 62$; 57.3/118.9). Their absolute magnitudes should not be compared across unequal sample sizes. The pinned MIDI-as-MTF loader (`midi_as_mtf_loader.py` in the FMD package) includes both the TPQ header and tick-valued message times. CLaMP-3 uses a related MTF front end; its 500 paired embeddings from the same lossless 220-to-440-TPQ intervention have Fréchet distance 30.55 and mean matched cosine 0.912. A fresh repeat extraction of the original 500 files has exactly zero feature-element difference. These are embedding sensitivities, not listener outcomes. The uncached feature arrays and file hashes are in `eval/tempo_metadata_control/clamp3_tick_invariance/report.json`. The pinned 64-character M3 patcher changes patch count for 144/500 twofold and 236/500 fourfold copies; 1/500 and 4/500 also cross a 512-patch segment boundary. These counts are reproduced by `pmt/audit_clamp2_patch_units.py`; the intervention does not isolate header, numerical-string and patch-boundary effects.

For bootstrap draw $b$, we sample 100 shared prompt indices with replacement and independently resample the same 500 real indices for every arm. The saved CLaMP-2 loader-order map binds

| Source | $n$ | 220 vs 440 TPQ | 220 vs 880 TPQ |
|---|---|---|---|
| All sources | 500 | **35.4** | 78.1 |
| Folk | 300 | **40.7** | 84.1 |
| Balanced | 132 | **50.1** | 93.5 |
| Classical | 62 | **57.3** | 118.9 |

Table 8: CLaMP-2 FMD between each real subset and its exact tick-rescaled copy, with unchanged note seconds and MIDI messages. Jazz ($n = 6$) is too small for a separate estimate; subset FMD magnitudes also depend on sample size.

| Serializer | Raw FMD | Strict 220 TPQ | Lossless common 10,560 TPQ |
|---|---|---|---|
| PMT | 149.0 | 149.0 | 107.4 |
| PerTok | **140.3** | 128.6 | **98.6** |
| Beat-TSD | 285.4 | 163.4 | 128.5 |
| REMI | 276.1 | 128.5 | 106.8 |
| MIDI-Like | 287.4 | 136.9 | 120.3 |
| Structured | 282.1 | **122.4** | 98.8 |
| ABC | 351.1 | 285.7 | 305.6 |

Table 9: The same seven matched 0.8B generation sets and clean 500-piece real reference. Strict 220 TPQ changes generated tick units while retaining all messages and tempo maps. At common 10,560 TPQ, real and generated files are transformed by integral factors, with every note time and non-time message unchanged. The very fine common unit is a sensitivity, not an endorsed CLaMP-2 operating point.

each feature row to its MIDI SHA256 and prompt index; array rows are not assumed to follow sorted filenames. Invalid generated files are excluded after prompt resampling. Let $\Delta^*_{jb}$ be the FMD difference for competitor $j$, $\bar{\Delta}^*_j$ and $s^*_j$ its bootstrap mean and standard deviation, and $\hat{\Delta}_j$ the full-sample difference. We compute

$$T_b = \max_{j=1}^{6} \left| \frac{\Delta^*_{jb} - \bar{\Delta}^*_j}{s^*_j} \right|, \qquad I_j = \left[ \hat{\Delta}_j - q_{.95}(T) s^*_j,\ \hat{\Delta}_j + q_{.95}(T) s^*_j \right]. \tag{8}$$

The 1,000-draw critical values are 2.414 at 220 TPQ and 2.481 at 10,560 TPQ (fixed seeds 20271146 and 20281486). These are simultaneous sampling intervals under the chosen reference and trained checkpoints, not confidence statements about listener quality or retraining.

## H PMT Encoding and Restricted Round Trip

The encoder sorts notes by absolute onset, track and pitch, rounds each absolute onset to the 10 ms lattice, and encodes positive differences with one or more shift tokens. It emits no shift for simultaneous notes. A track/program pair appears when the active track changes, then each note emits pitch, clipped duration and a 32-level velocity. Decoding adds the shift values to an absolute tick accumulator. For at most 16 input instrument entries (including empty tracks), valid program and pitch IDs, and durations in the unclipped 10–2000 ms interval, pitch, non-drum program, drum flag and grouping by nonempty instrument track round-trip, but track indices may be renumbered by first-onset order. Quantized onset/duration error is at most 5 ms in memory. Export to a 220-TPQ, 120-BPM MIDI file adds up to 1.14 ms rounding per event if note identity survives; onset error can then reach 6.14 ms, and duration error can reach 7.28 ms after separate endpoint rounding. A MIDI writer may also resolve overlapping notes of the same pitch, so the in-memory note-identity guarantee does not automatically transfer to a written file. The original program number of a drum track is not retained because the decoder assigns its program to zero. Longer durations can have arbitrarily large clipping error. The guarantee concerns the serializer, not generated note validity or acoustic similarity.

| Contrast | Strict 220 TPQ | Exact common 10,560 TPQ |
|---|---|---|
| Beat-TSD–PMT | 14.4 [-1.6,30.4] | 21.1 [7.2,34.9] |
| PerTok–PMT | -20.4 [-40.3,-0.6] | **-8.9 [-28.4,10.7]** |
| REMI–PMT | -20.5 [-39.7,-1.4] | -0.6 [-21.8,20.5] |
| MIDI-Like–PMT | -12.2 [-31.4,7.1] | 12.9 [-8.6,34.4] |
| Structured–PMT | **-26.6 [-45.1,-8.1]** | -8.7 [-29.0,11.6] |
| ABC–PMT | 136.7 [104.0,169.4] | 198.1 [153.5,242.8] |

Table 10: All six competitor-minus-PMT FMD point contrasts and simultaneous 95% bootstrap intervals. Both protocols use the same seven frozen generation sets and 500 real scores; interval widths exclude retraining variability.

## I Corpus Construction and 6.25M Checkpoints

The 86,598-case corpus links caption, MIDI, ABC and rendered audio at piece level. We parse and normalize source MIDIs, freeze a source-stratified split, render excerpts with FluidSynth, ask Qwen3-Omni for audio-grounded descriptions with score metadata, and filter text for language, repetition, refusal and musical content. Native ABC may contain more of a folk tune than the rendered MIDI slice; the controlled ABC cache uses midi2abc on the same slice as all other arms. The larger manifest has 6,250,143 prefilter pairs and 6,248,214 retained rows. Its published statistics list 6,244,793 unique hash IDs; path-level deduplication did not eliminate every repeated hash. The documented held-out-content index covers the 86k test set, not the MidiCaps sample. Captions have not been independently verified by human listeners; raw redistribution depends on each source license.

**Scope relative to earlier resources and models.** Symbolic corpora cover different domains: MAESTRO and GiantMIDI-Piano emphasize piano performance (Hawthorne et al., 2019; Kong et al., 2020); POP909 emphasizes pop arrangements (Wang et al., 2020); MetaMIDI links MIDI with audio metadata (Ens & Pasquier, 2021); Aria-MIDI and GigaMIDI broaden the scale of piano and general MIDI collections (Bradshaw & Colton, 2025; Lee et al., 2025). Their sizes are not interchangeable with the number of caption-aligned, filtered training slices here. MusicBERT studies symbolic understanding (Zeng et al., 2021); MMM, FIGARO and MuPT investigate different conditional generation or ABC modeling recipes (Ens & Pasquier, 2020; von Rütte et al., 2023; Qu et al., 2025). We do not count those systems as controlled tokenizer arms without reimplementing their data and training recipes.

The PMT-4B initial 166k-step checkpoint is `pmt6M_ckpt166000`; the evaluated continuation is `q35_4B_pmt_6M_ep2/checkpoint-166000`. The latter was trained for 166k additional optimizer steps. Both contain readable 8,413,432,320-byte safetensors files; the nominal "epoch" labels are only step-schedule names because the streaming loader does not log complete data passes. Against the self-built reference, FMD moves 144.8 to 152.5, while against the old source-stratified reference it moves 123.5 to 117.0 on the same 200 captions. On the 165 original-MIDI MidiCaps-source reference, FMD moves 351.0 to 313.5; corrected exact-key accuracy moves .139 to .182 and coarse instrument-family F1 .597 to .627. Yet 159/165 sources match the 6M training manifest by exact quantized-note fingerprint. This probe is not held out and its checkpoint changes cannot establish generalization. The old unrelated piano-collapsed 200-piece reference gives 210.95 and 224.91; independent CLaMP-2 feature extraction reproduces these legacy FMD values, but they are not a caption-aligned benchmark. Both endpoints export 120-BPM MIDI. The full paths and audit are stored in `ICLR2027/6M_checkpoint_audit.md`.

## J External Baseline and Perceptual Scope

Table 11 uses 165 captions and their original MIDI files. It excludes 35 IDs from the 86k controlled training set but retains 159 exact quantized-note matches to the 6M stream, making the 6M rows training-exposed. The published systems differ in model, representation, data and training procedure; their rows diagnose metrics and attributes rather than provide a causal or held-out system comparison. The old 200-piece FMD reference was constructed from different examples after note truncation and

| System | FMD↓ | Exact key↑ | GM-family F1↑ |
|---|---|---|---|
| PMT 86.6k | 365.4 | 0.073 | 0.568 |
| PMT 6M, 166k | 351.0 | 0.139 | 0.597 |
| PMT 6M, +166k | 313.5 | 0.182 | 0.627 |
| MIDI-LLM | 298.4 | 0.303 | 0.806 |
| MIDILM | 272.6 | 0.521 | 0.737 |
| Amadeus | **223.9** | **0.618** | **0.808** |
| text2midi | 420.2 | 0.248 | 0.548 |
| ChatMusician | 414.1 | 0.333 | 0.293 |

Table 11: Training-exposed diagnostic on 165 custom MidiCaps-source captions after excluding 35 source IDs present in the controlled 86k training corpus; 159/165 still match the 6M training manifest by quantized-note fingerprint. FMD uses CLaMP-2 on 165 saved generations against the corresponding original MIDIs. Exact key compares pitch class and mode after music21 estimation; F1 uses canonical coarse GM families on 154 captions naming an instrument. Three published-system directories lack independently verified full prompt contents. These rows do not measure held-out 6M performance or define a controlled system ranking.

piano collapse, so those values are retained only as an archival diagnostic. All PMT 6M outputs here encode a 120-BPM default tempo. The archived decode-time constraint pilot used flawed adherence extractors and is not quoted numerically. A prior human-pilot summary has no recoverable response log in this workspace and supplies no result.

### J.1 Saved Outputs and Source-Aligned References

We evaluate archived outputs from PMT and five published text-to-MIDI systems, including Amadeus (Su et al., 2025) and MIDILM (Li et al., 2026), to diagnose metric and caption-control failure modes. The original 200-caption file is a *custom MidiCaps sample*, not its official test split: only 13 rows have `test_set=true`, and 35 source MIDI IDs occur in our 84,576-piece training set. We show the 165 remaining indices with their original, caption-aligned real MIDIs as an FMD reference. A second audit, using the larger training manifest's exact quantized-note fingerprint, finds *194/200* custom items in the 6.25M training stream, including *159/165* of these rows. Thus the PMT-6M entries are training-exposed diagnostics, not generalization or system-superiority evidence. Surviving generation logs preserve indices 0–199 but do not independently prove exact caption contents for three published-system directories; systems also differ in model, corpus and decoder.

On this unmodified-reference subset, PMT's 6M continuation changes FMD from 351.0 to 313.5, exact-key agreement from 0.139 to 0.182, and GM-family F1 from 0.597 to 0.627. These checkpoint changes occur on a set with 159/165 training-exposed targets and do not measure generalization. The historical 200-piece FMD reference in `eval/t2m_ref` instead draws unrelated samples, truncates each to 256 onset-ordered notes and collapses every instrument and drum into one piano track. Its earlier FMD values (211 and 225 for the two 6M endpoints) measure that transformed distribution and must not be compared with Table 11. Both PMT 6M endpoints export default 120-BPM tempo, so tempo-bin matches do not show prompt following. Native key control remains weak relative to Amadeus; a chord-sequence and requested-duration evaluation is still needed. An archived 80-prompt decode-time constraint pilot used the flawed old key/keyword extractors, so its numerical improvement is withheld pending a corrected re-score.

**Reference-choice diagnosis.** PMT 6M +166k has FMD 313.5 on full original scores versus Amadeus 223.9, but 177.4 on 256-note prefixes versus Amadeus 297.7 (Figure 4). Every system keeps exactly the same 165 generated files; only the real feature distribution changes. Training exposure prevents a generalization comparison here, but cannot explain a within-output reversal caused solely by the reference. This is a concrete failure mode for a single unnamed FMD music-quality ranking. We therefore treat the clean, tune-disjoint 500-piece comparison in Table 1 as a specified distribution-matching task, report duration and validity alongside it, and avoid inferring perceptual superiority from an FMD difference. The reference intervention also changes duration and tempo information, so it is a sensitivity test rather than an estimate of metric error.

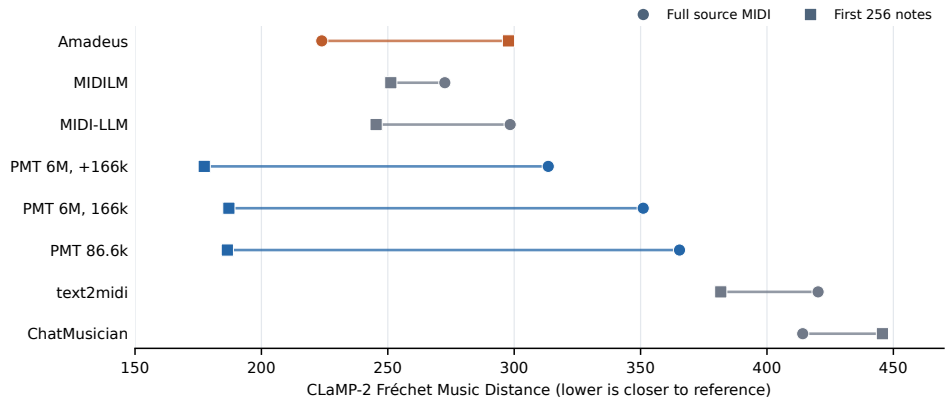


Figure 4: Reference sensitivity on the *same* saved generations and 165 custom caption indices, 159 of which match the PMT-6M training manifest. Circles use full original caption-aligned MIDIs; squares use the first 256 onset-ordered notes, retaining instrument programs and drum flags. The second reference shifts onset zero and keeps the initial tempo; length and later tempo changes differ. The PMT/Amadeus ranking reverses without changing a model or generated file. This diagnostic does not measure held-out music quality.

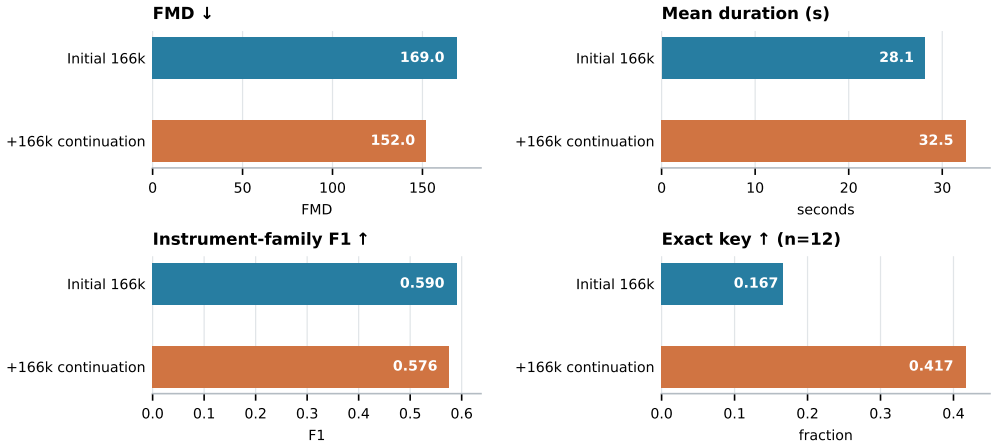


Figure 5: Two saved PMT-4B 6M checkpoint endpoints on the same 100 captions and fixed 500-piece reference. The continuation is 166k additional updates; both outputs are 100/100 valid. Instrument F1 falls slightly while FMD decreases, and exact-key rate uses only 12 explicit-key prompts. These single-run endpoints do not demonstrate a monotonic or held-out gain.

### J.2 Large-Corpus Checkpoint Audit

Two PMT-4B checkpoints, 166,000 updates apart, generated on the same 100 captions with a 640-token cap. The 500-piece reference has zero exact quantized-note fingerprint matches to the 6,248,214-row training manifest; near duplicates remain possible.

Initial/continued FMD is 169.0/152.0, with 100/100 valid outputs; mean duration is 28.1/32.5 seconds. Instrument-family F1 is 0.590/0.576, and exact-key match is 0.167/0.417 on only 12 captions. The paired initial-minus-continued FMD difference is 17.0 (95% bootstrap interval [-2.7,31.2]), spanning zero. This same-tokenizer comparison cannot identify a representation effect; the custom MidiCaps probe is training-exposed (Appendix J).

## K Canonical Note-Only Evaluation

The converter uses integer microseconds before rounding to the nearest millisecond, avoiding floating-point half-millisecond ties across equivalent TPQ exports. It retains track order, program, drum flag, pitch, velocity, onset and offset, and sorts notes deterministically. Other MIDI events are dropped. All 500 real and 693 valid generated outputs match their exact tick-rescaled counterparts byte for byte after canonicalization. Features were re-extracted uncached; row maps store each MIDI SHA256.

**Event scope of this audit.** All 500 reference files and 693 valid generated files have exactly one initial tempo event and a 4/4 signature. No file contains a controller, pitch bend, aftertouch, polyphonic aftertouch or system-exclusive event. The generated set contains 98 ABC key signatures and descriptive text/track-name events, which the projection drops. The generated tempo headers comprise 396 files at 120 BPM and 297 at approximately 121.29 BPM; the projection preserves the decoded note times in seconds while replacing those headers. The zero null therefore establishes invariance for the audited note content, not for expressive controller streams or alternate meters.

**Public generated-score replication.** The anonymous supplement includes 21 selected generated MIDIs (not the 500 real reference scores), each with exact $2\times/4\times$ tick-unit copies. A standalone script verifies all message fields, delta-tick scaling and note seconds; uncached CLaMP-2 extraction gives FMD 73.0/94.3 between originals and their copies. The small-sample magnitudes are not comparable to the 500-piece null, but they allow a public end-to-end check of the failure mode.

| Serializer | Valid | Canonical FMD |
|---|---|---|
| PMT | 98 | 151.6 |
| PerTok | **100** | 131.1 |
| Beat-TSD | 98 | 189.4 |
| REMI | **100** | 143.0 |
| MIDI-Like | 99 | 151.4 |
| Structured | **100** | **126.3** |
| ABC | 98 | 345.4 |

Table 12: Canonical note-only CLaMP-2 FMD against the same canonical clean500 reference.

## L Length and Decode Budget

The 900-token generation cap is the same for every serializer, but it does not imply the same musical duration: token cost per note and note density vary. PMT has no learned requested-duration field. A separate paired test on the 6.25M PMT-4B checkpoint generated 100 captions with 900 and 7,000 new-token caps. Median note count rose from 195.5 to 704.5 and median duration from 18.25 to 44.27 seconds; 90/100 captions gained notes and 81/100 gained duration. The corresponding FMD against a clipped reference changed from 231.2 to 299.6, whereas truncating each long generation to its paired reference note count gave 229.3 (source artifacts: `eval/longgen_compare.json` and `eval/longgen_fmd.json`). These are different output-length protocols, not evidence that the longer pieces sound worse. They show why the primary table includes duration and why a requested-length interface and length-matched evaluation should accompany future music-quality claims.

## M Archived Qualitative Renderings

These selected plots and engravings are restored as qualitative records from the earlier manuscript. They draw on mixed checkpoints and source domains; they are neither a random sample nor a matched evaluation of the new seven-arm grid. Complete caption bindings for some published-system output directories cannot be independently verified. A piano roll and an engraved excerpt show what a saved MIDI contains; engraving quantizes notes for legibility. These panels do not establish listener preference, full-caption control, or the current FMD ranking.

### M.1 Generated Piano Rolls and Scores

Each plate contains the printed caption and model label, a piano roll, and an engraved opening. The staff programs are the generated MIDI programs, so caption–instrument mismatches remain visible.

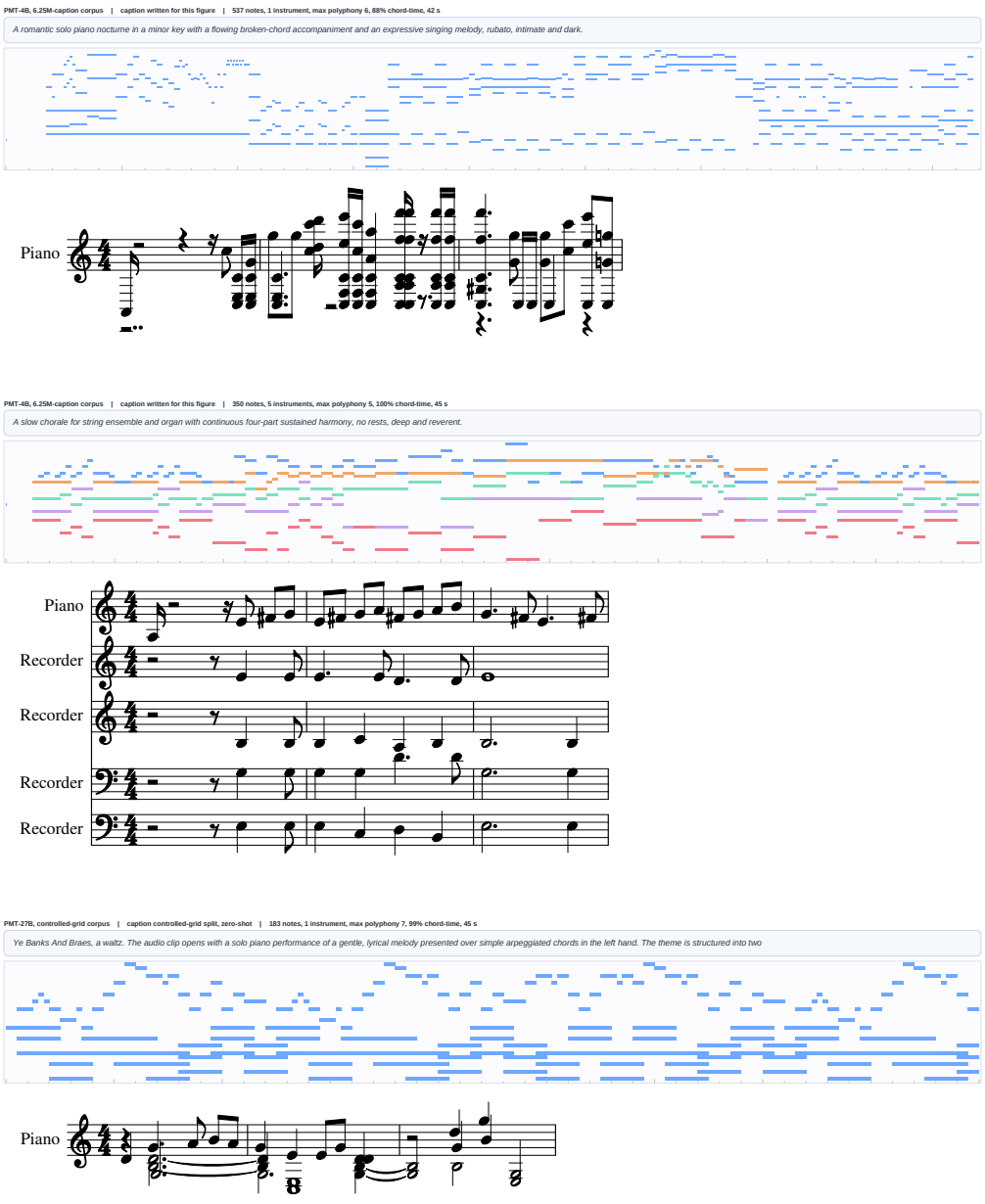


Figure 6: Selected archived generations, plate 1 of 11. Caption, piano roll, and engraved excerpt are views of the saved MIDI. The model label is printed on each panel. Examples are illustrative, not estimates of typical output.

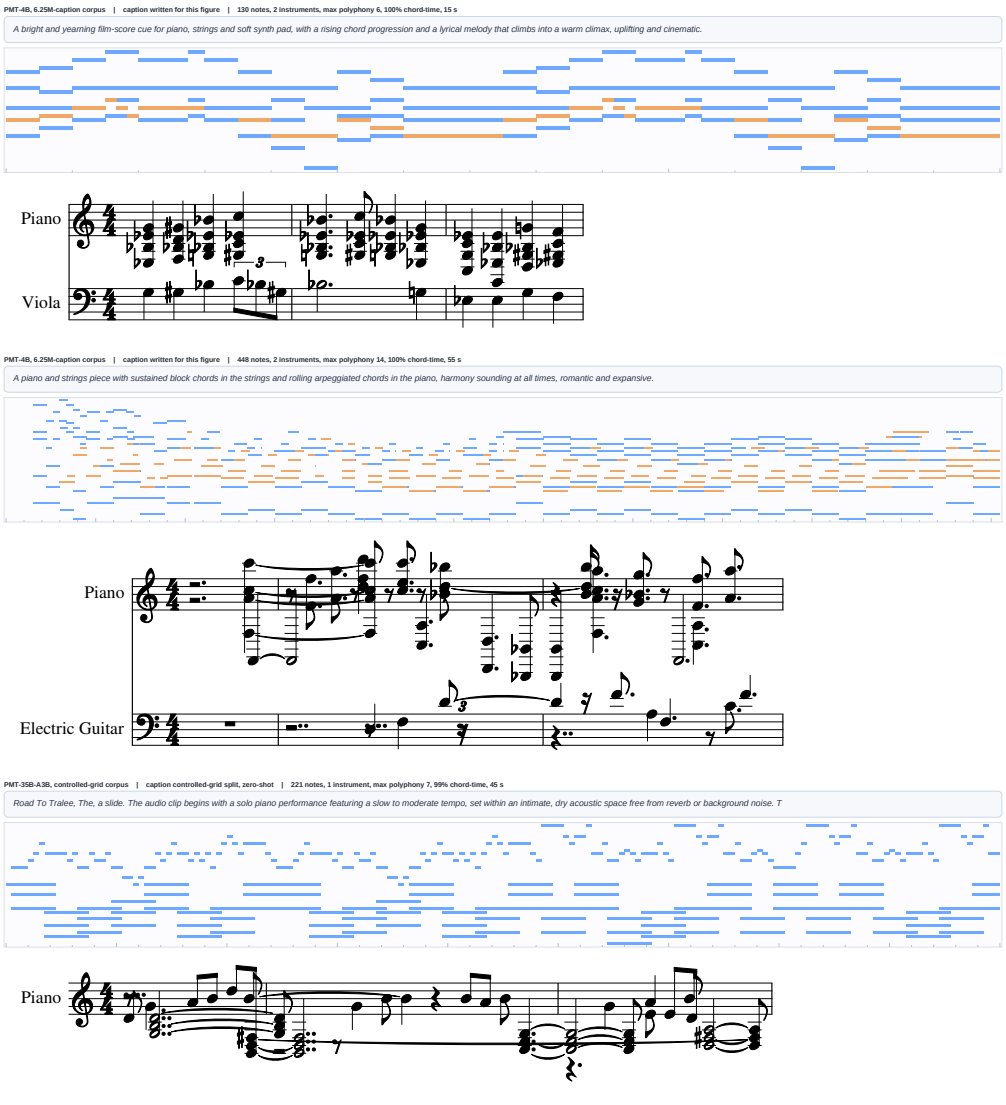


Figure 7: Selected archived generations, plate 2 of 11. Caption, piano roll, and engraved excerpt are views of the saved MIDI. The model label is printed on each panel. Examples are illustrative, not estimates of typical output.

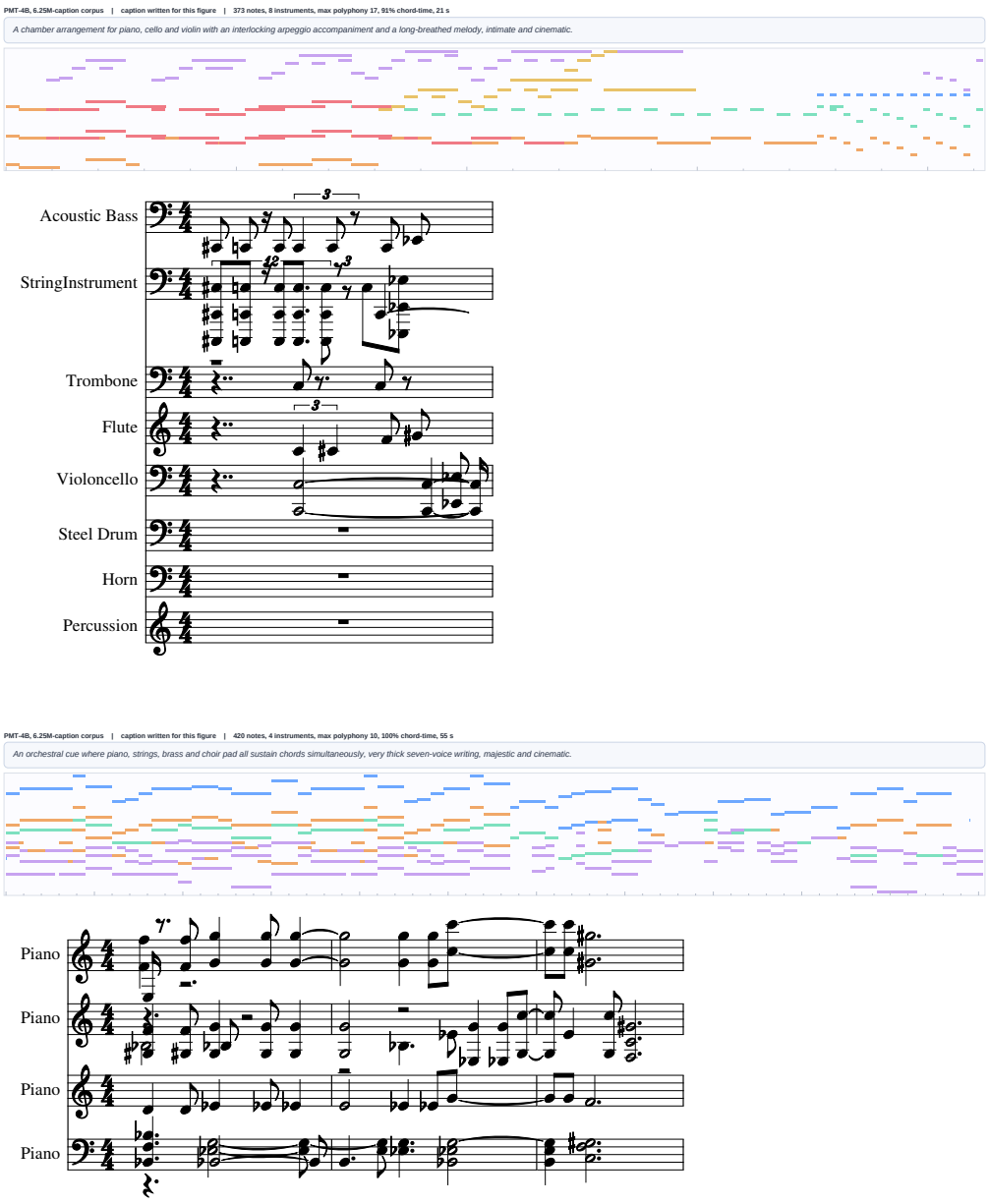


Figure 8: Selected archived generations, plate 3 of 11. Caption, piano roll, and engraved excerpt are views of the saved MIDI. The model label is printed on each panel. Examples are illustrative, not estimates of typical output.

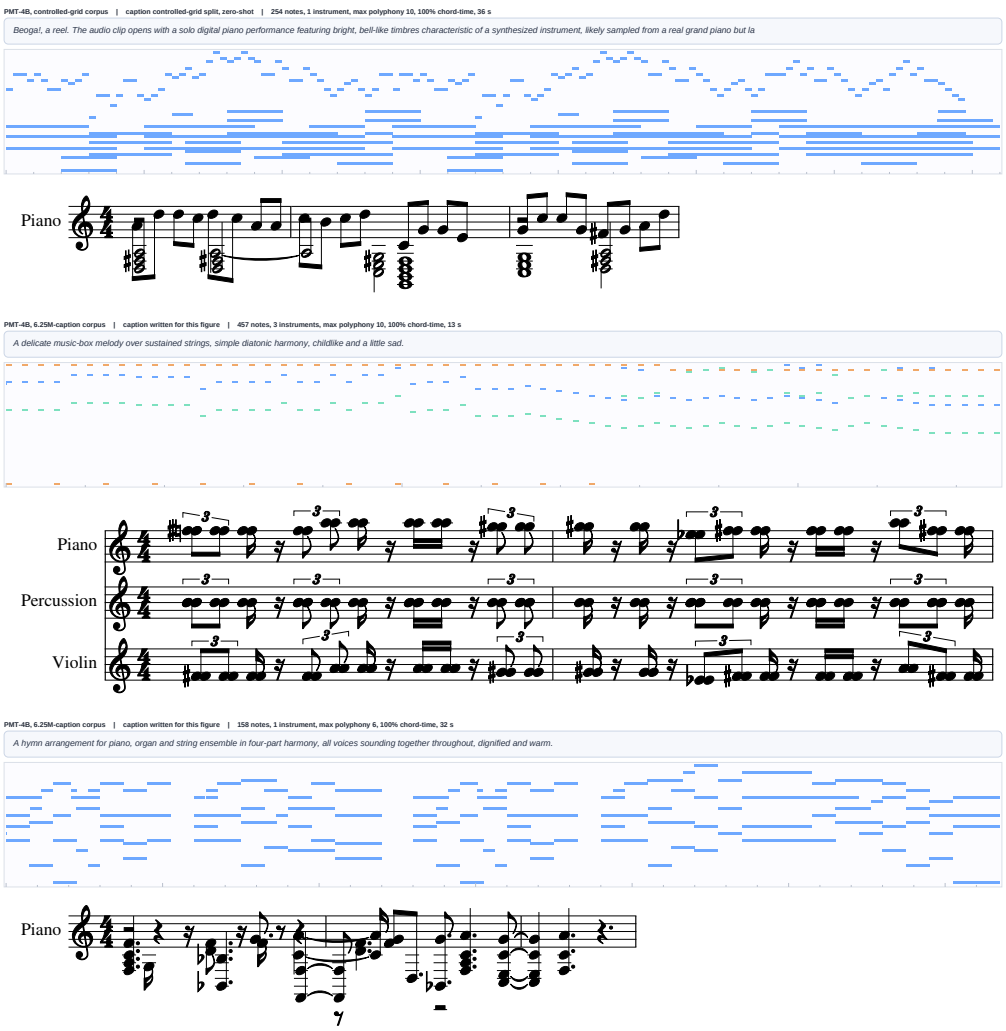


Figure 9: Selected archived generations, plate 4 of 11. Caption, piano roll, and engraved excerpt are views of the saved MIDI. The model label is printed on each panel. Examples are illustrative, not estimates of typical output.

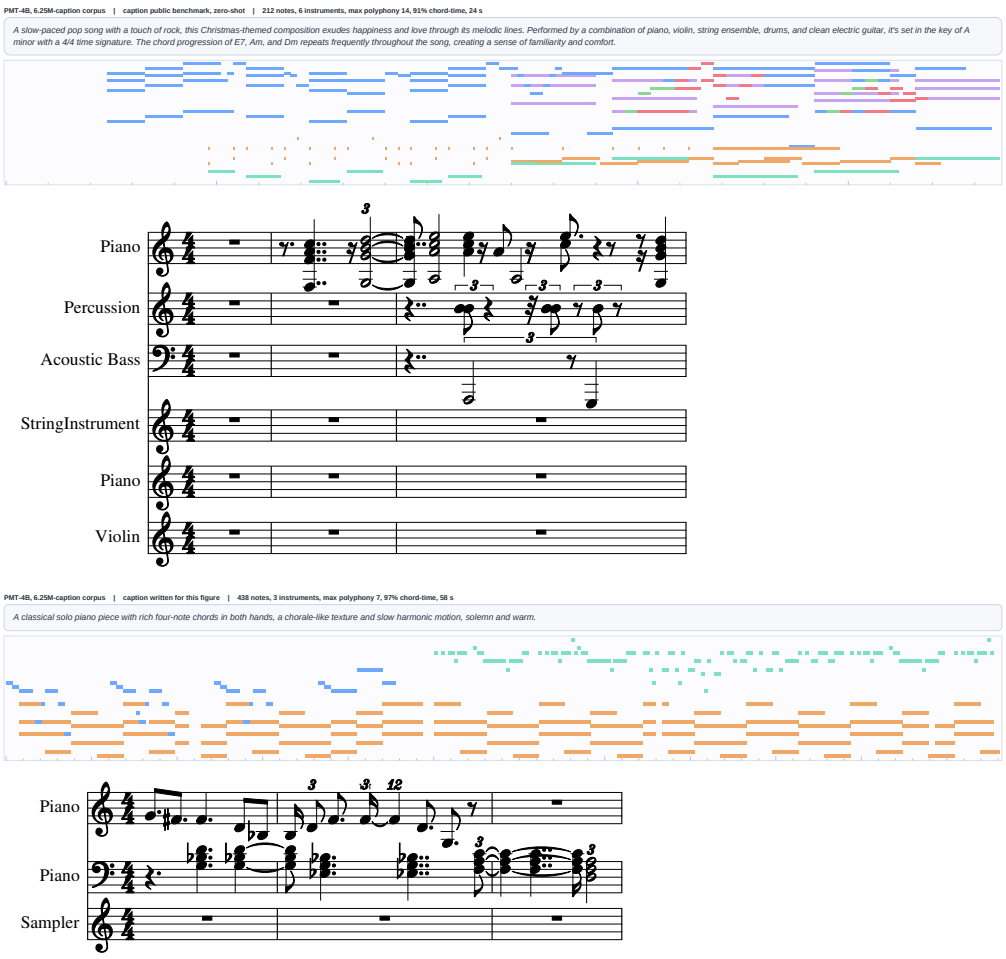


Figure 10: Selected archived generations, plate 5 of 11. Caption, piano roll, and engraved excerpt are views of the saved MIDI. The model label is printed on each panel. Examples are illustrative, not estimates of typical output.

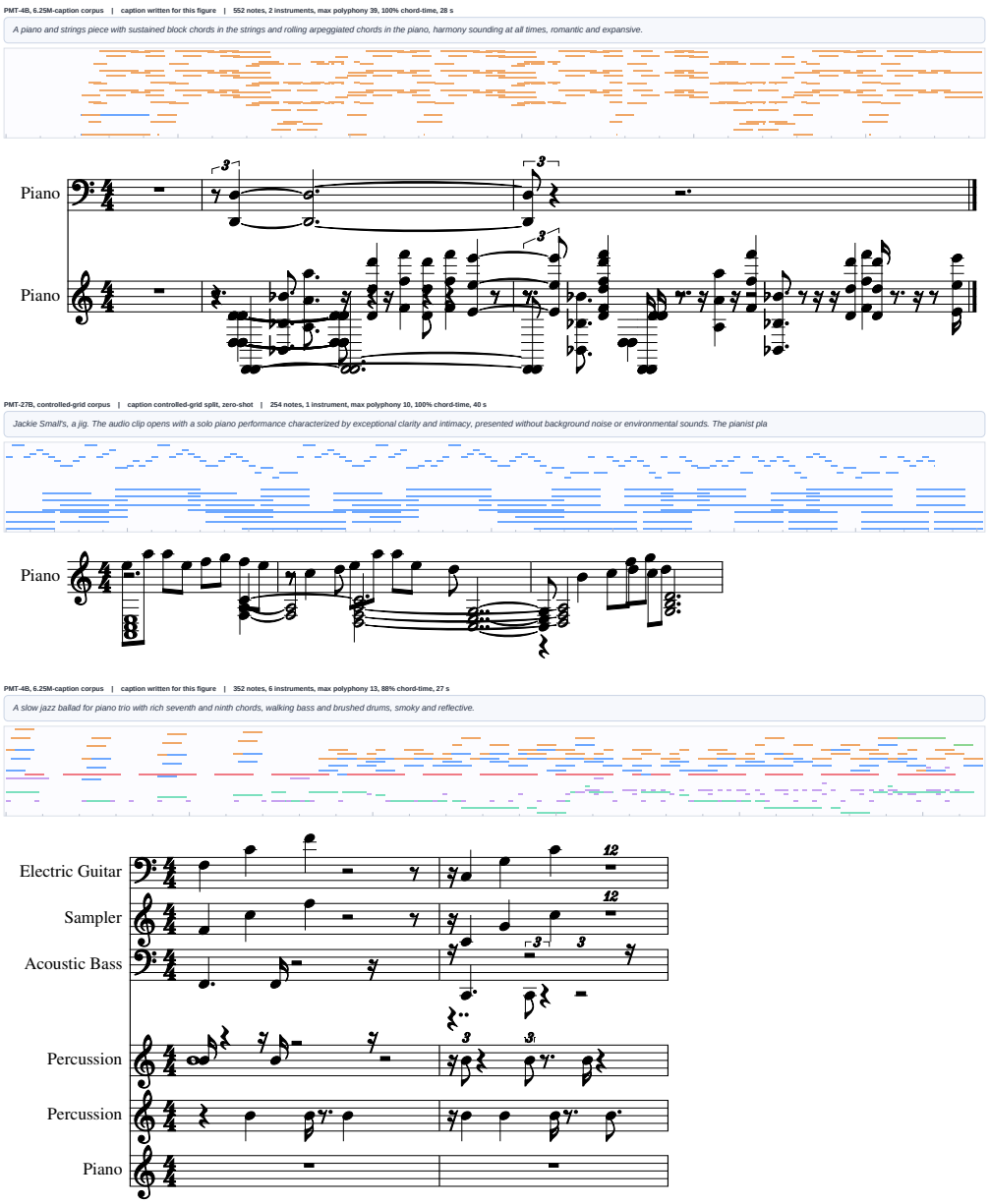


Figure 11: Selected archived generations, plate 6 of 11. Caption, piano roll, and engraved excerpt are views of the saved MIDI. The model label is printed on each panel. Examples are illustrative, not estimates of typical output.

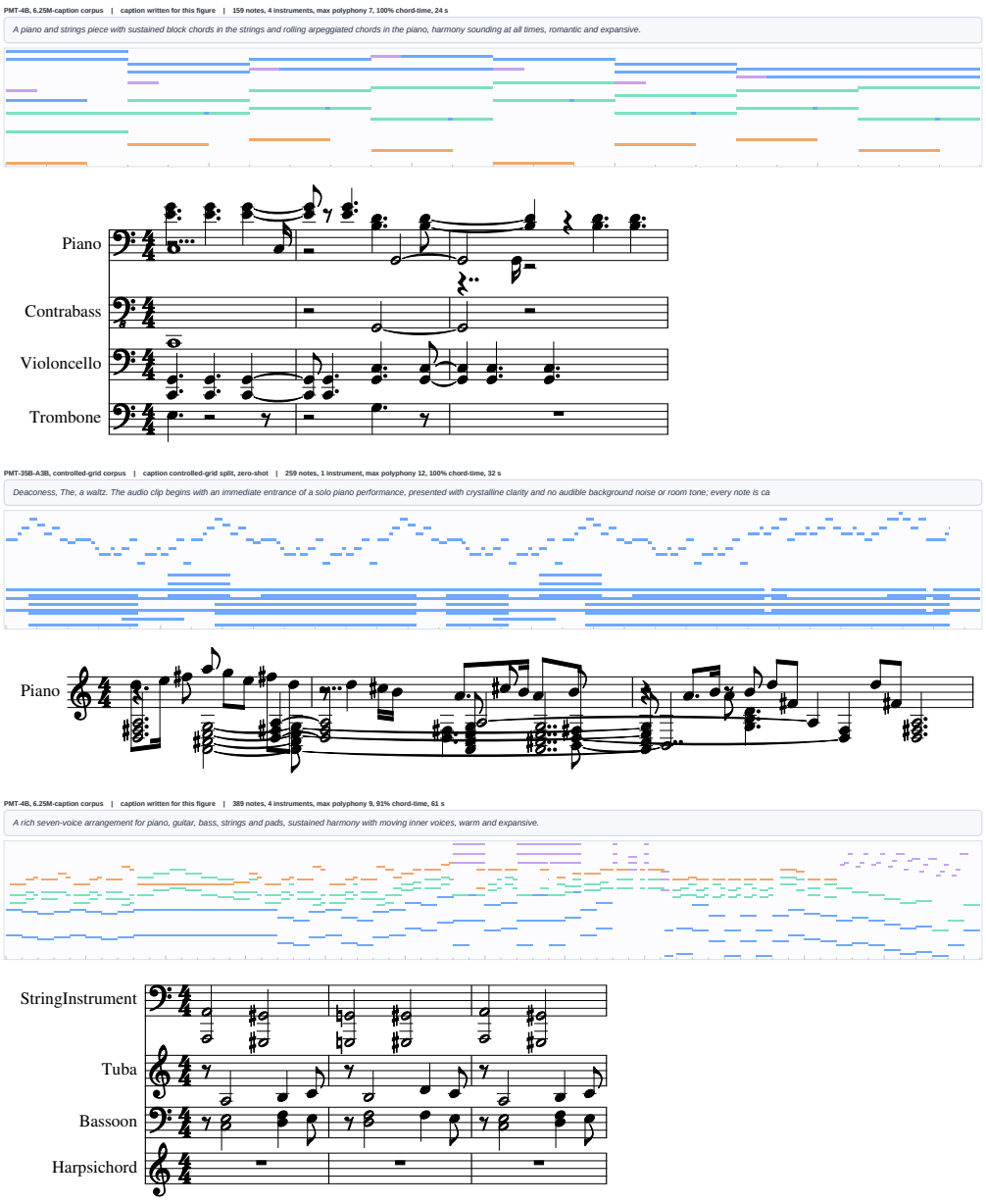


Figure 12: Selected archived generations, plate 7 of 11. Caption, piano roll, and engraved excerpt are views of the saved MIDI. The model label is printed on each panel. Examples are illustrative, not estimates of typical output.

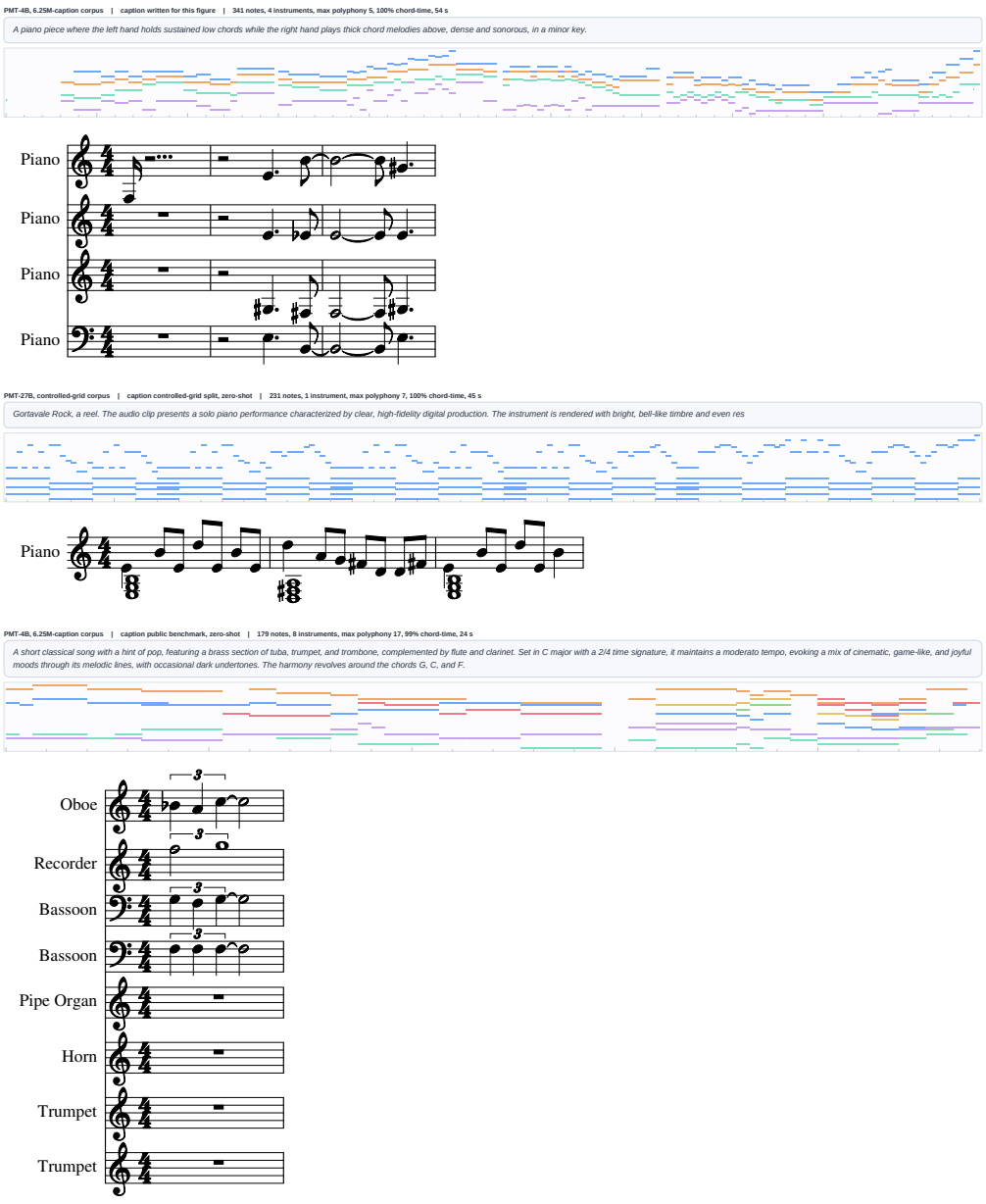


Figure 13: Selected archived generations, plate 8 of 11. Caption, piano roll, and engraved excerpt are views of the saved MIDI. The model label is printed on each panel. Examples are illustrative, not estimates of typical output.

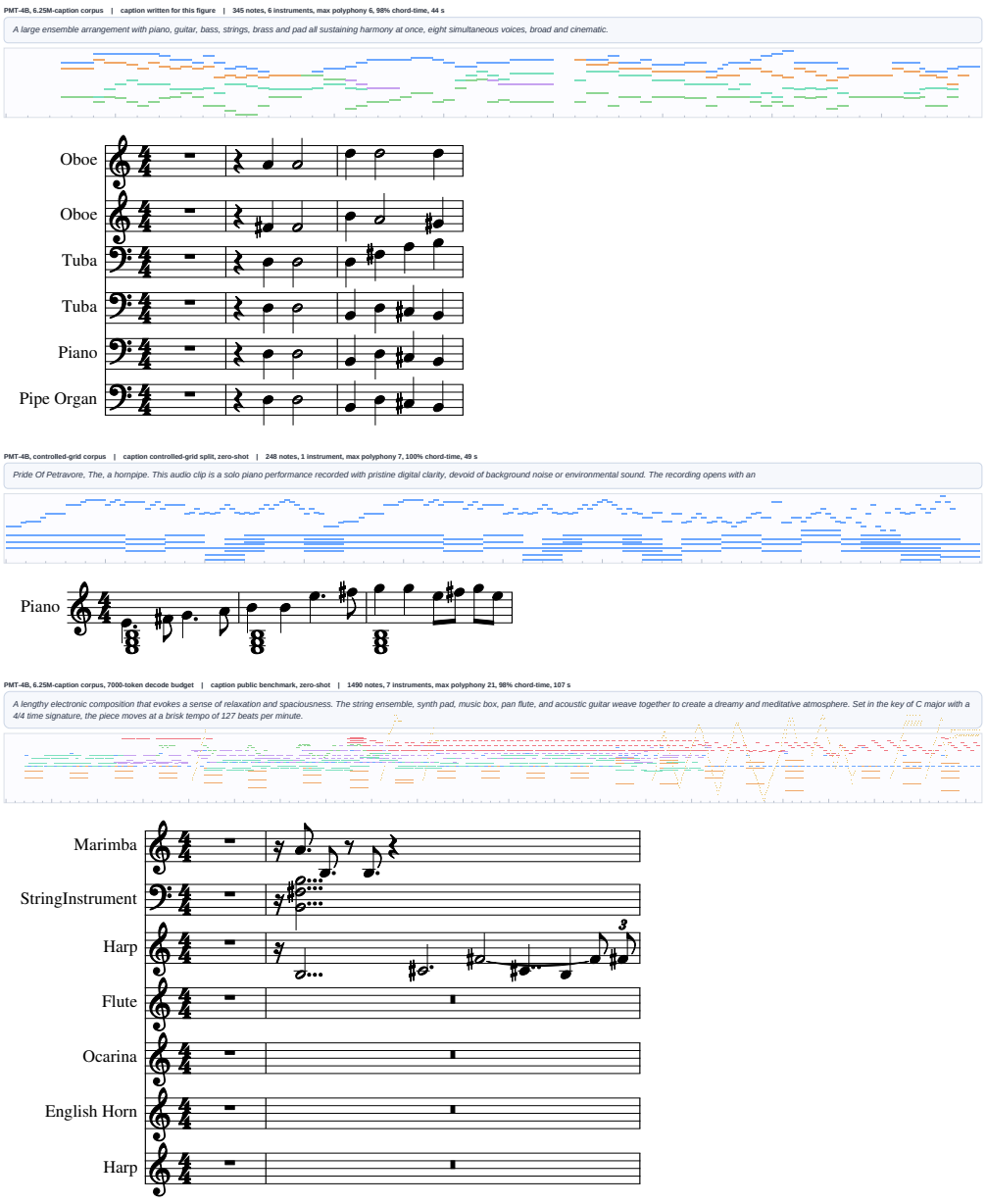


Figure 14: Selected archived generations, plate 9 of 11. Caption, piano roll, and engraved excerpt are views of the saved MIDI. The model label is printed on each panel. Examples are illustrative, not estimates of typical output.

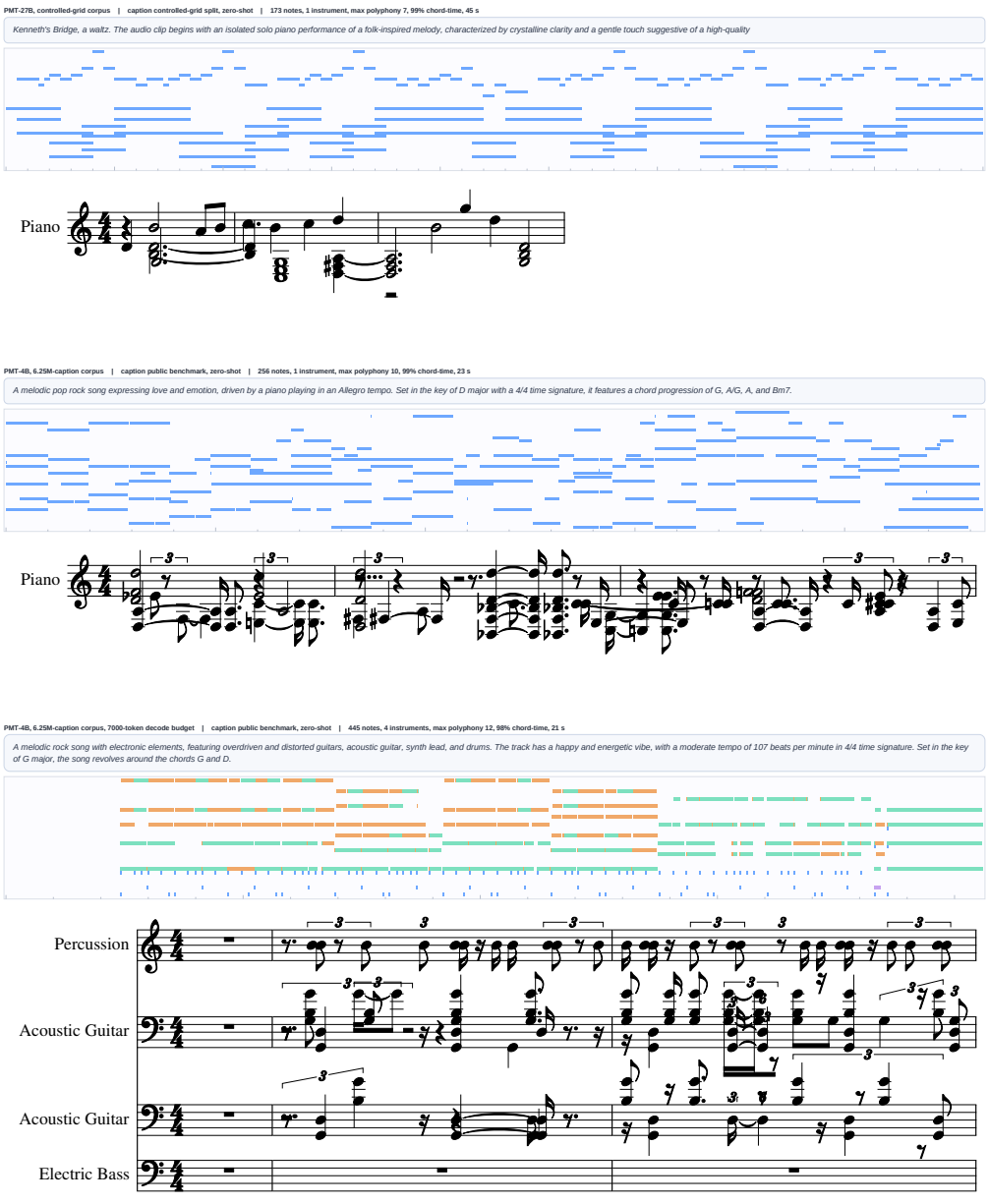


Figure 15: Selected archived generations, plate 10 of 11. Caption, piano roll, and engraved excerpt are views of the saved MIDI. The model label is printed on each panel. Examples are illustrative, not estimates of typical output.

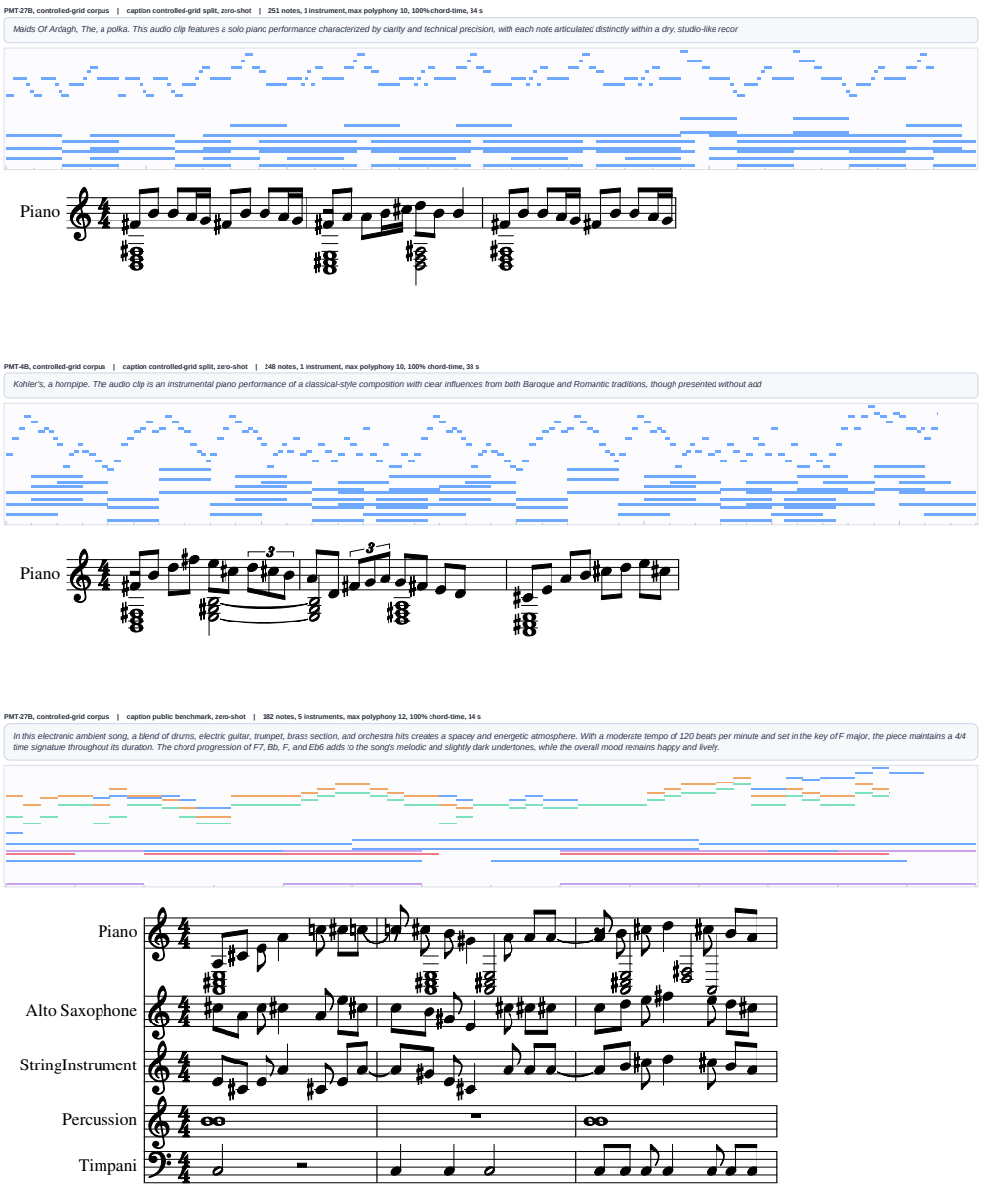


Figure 16: Selected archived generations, plate 11 of 11. Caption, piano roll, and engraved excerpt are views of the saved MIDI. The model label is printed on each panel. Examples are illustrative, not estimates of typical output.

### M.2 Same-Caption Comparisons

### M.3 Further Archived Visual Diagnostics

The following panels also use saved outputs from the earlier manuscript. Their quantitative interpretations are superseded by the audited reference and caption-cap protocol in the main paper; the panels are retained only to show the structure of those saved files.

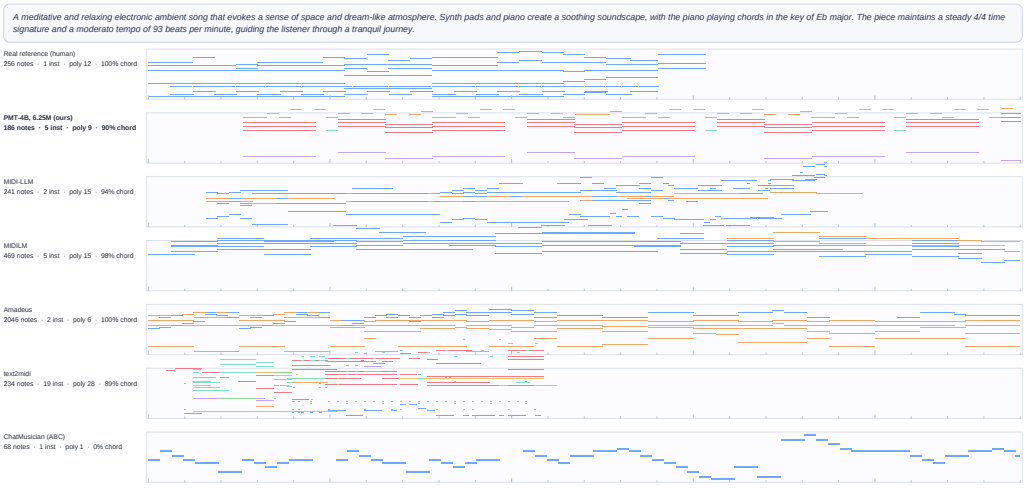


Figure 17: First archived caption: saved MIDI piano rolls from multiple systems over a common 24-second window. This selected example uses the older custom caption set, not a held-out 6M or controlled seven-serializer comparison. Appendix J audits the reference and training overlap.




Figure 18: Second archived caption: saved MIDI piano rolls from multiple systems over a common 24-second window. This selected example uses the older custom caption set, not a held-out 6M or controlled seven-serializer comparison. Appendix J audits the reference and training overlap.

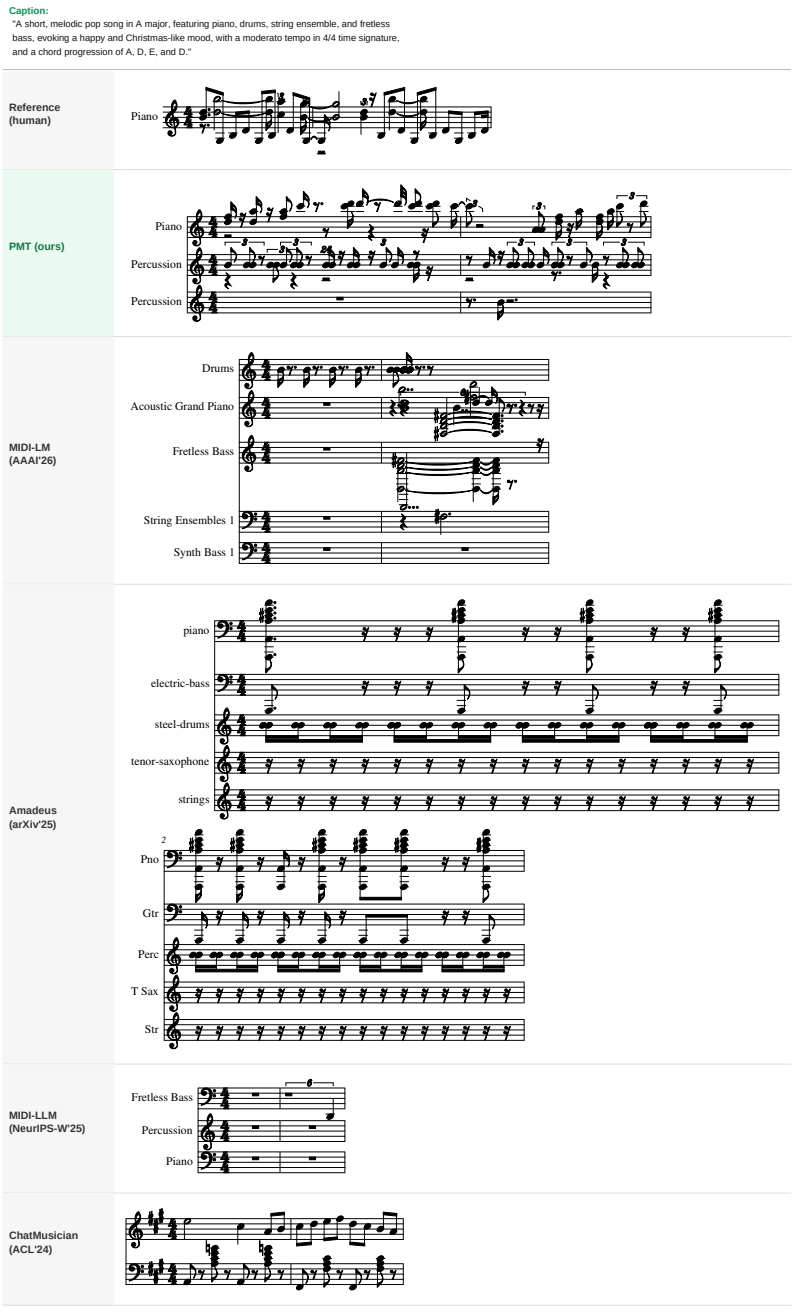


Figure 19: Pop caption: archived reference and generated excerpts engraved from saved MIDI after notation quantization. Staff labels show output MIDI programs. This single prompt does not compare systems under matched training, export, or listening conditions.

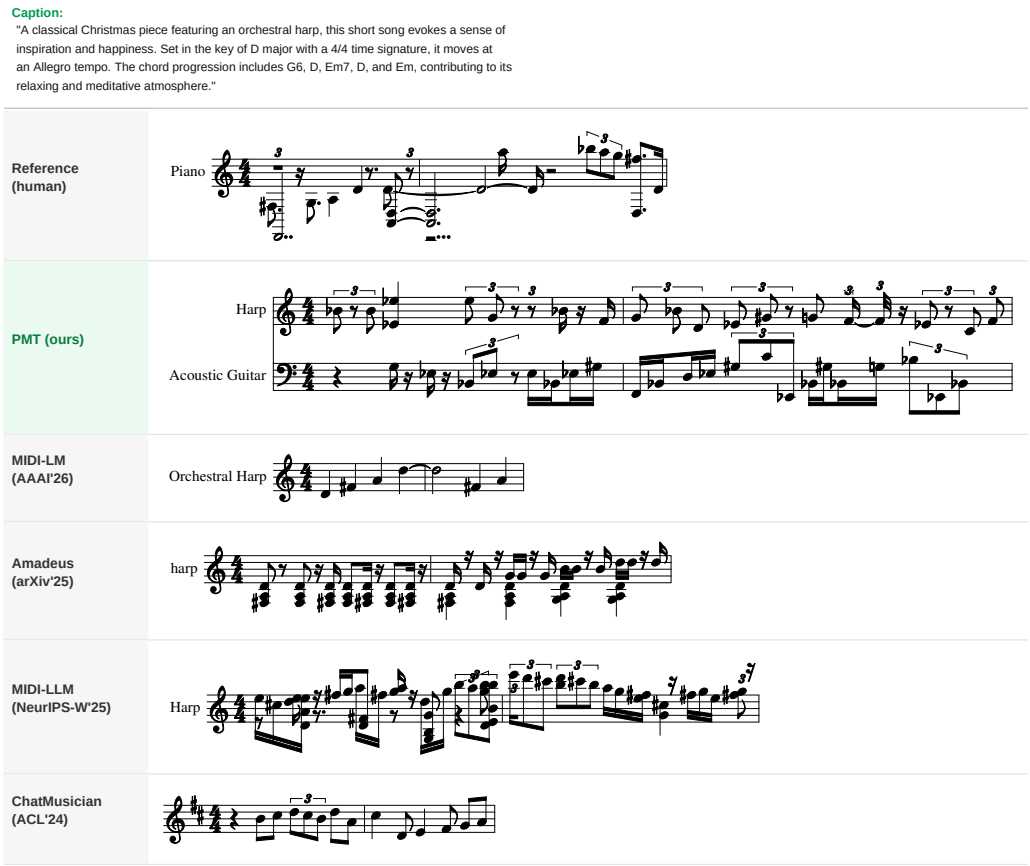


Figure 20: Harp caption: archived reference and generated excerpts engraved from saved MIDI after notation quantization. Staff labels show output MIDI programs. This single prompt does not compare systems under matched training, export, or listening conditions.

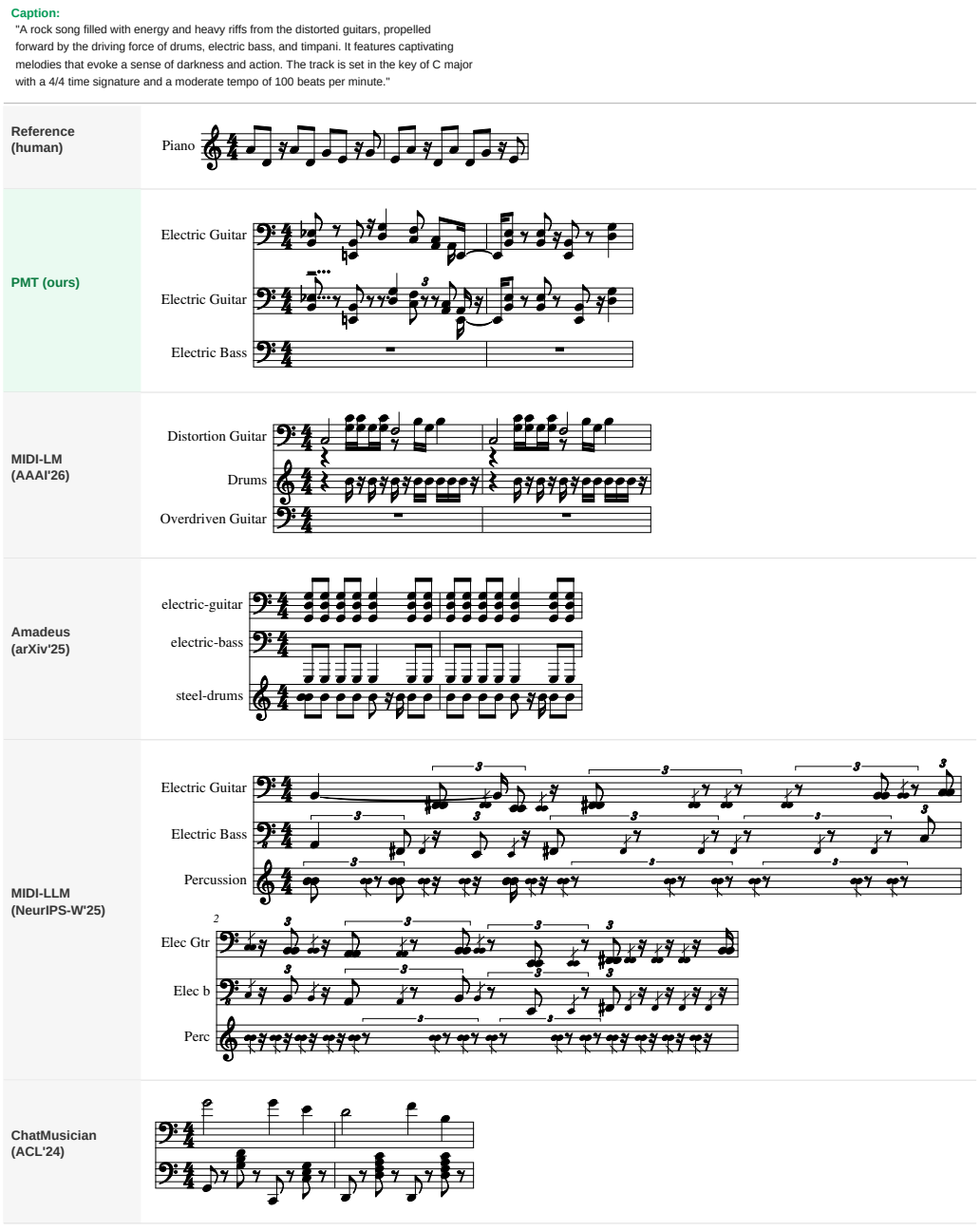


Figure 21: Rock caption: archived reference and generated excerpts engraved from saved MIDI after notation quantization. Staff labels show output MIDI programs. This single prompt does not compare systems under matched training, export, or listening conditions.

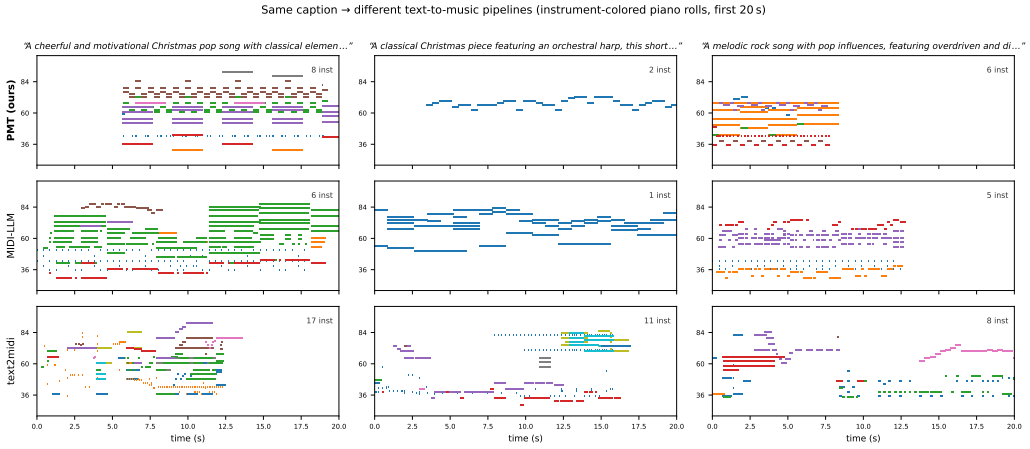


Figure 22: Three selected captions and archived system piano rolls. The outputs come from different training data and decoding pipelines. Complete caption bindings for some published-system directories are unverified, so the panel does not establish prompt adherence or system superiority.

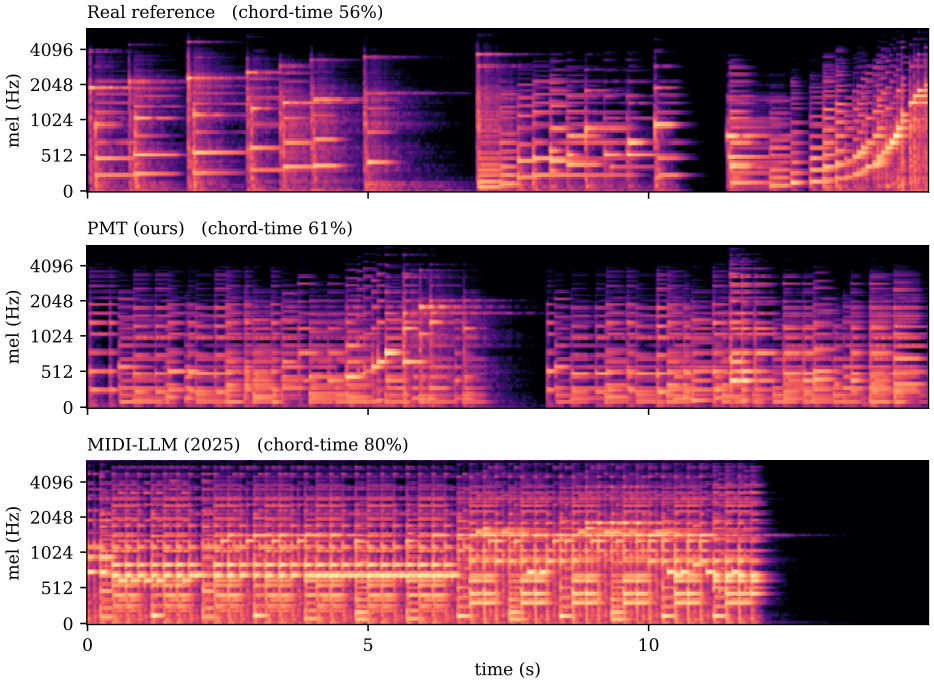


Figure 23: Archived soundfont renders and mel spectrograms for one selected caption. The same render settings were applied to the saved MIDI files; spectral density in one example is illustrative and is not a listening score.

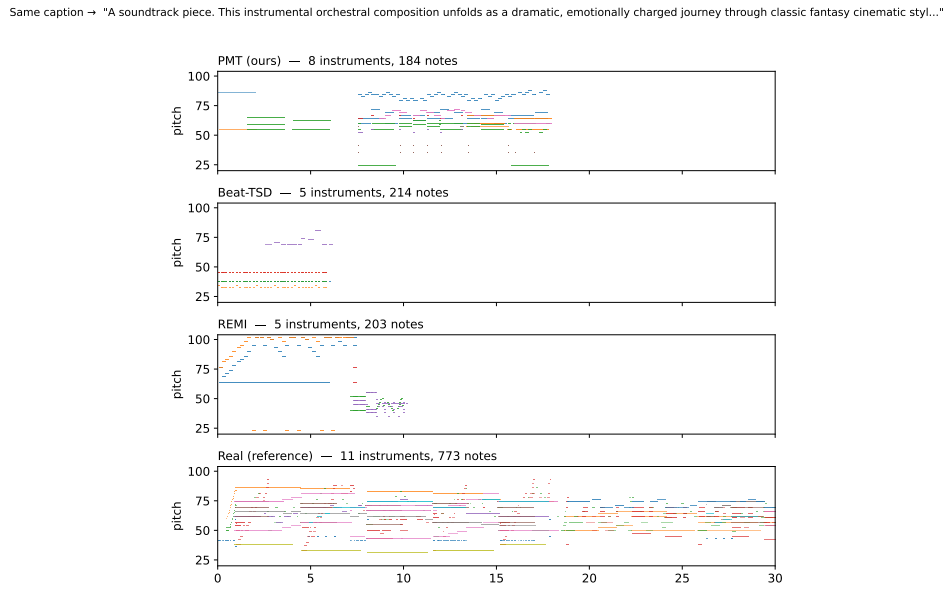


Figure 24: Archived same-caption piano rolls from the older tokenizer grid. This grid used unequal caption caps and a different FMD reference; the figure is retained as a MIDI-rendering example and does not support the controlled seven-arm conclusion of Table 1.