TS-RaMIA: Membership Inference Attacks for Symbolic Music Generation Models

Yuxuan Liu and **Rui Sang** and **Peihong Zhang** and **Zhixin Li** and **Kunyang Zhang** **Shengyuan He** and **Ye Li** and **Kaiyi Xu** and **Shengchen Li**
Xi'an Jiaotong-Liverpool University, Suzhou, China
{YUXUAN.LIU2204, RUI.SANG22, PEIHONG.ZHANG20, ZHIXIN.LI22, KUNYANG.ZHANG23, SHENGYUAN.HE23, YE.LI23, KAIYI.XU23}@STUDENT.XJTLU.EDU.CN
SHENGCHEN.LI@XJTLU.EDU.CN

Abstract

Artists and rights holders face growing concerns about unauthorized use of their copyrighted works in training generative models. We introduce **TS-RaMIA**, a practical auditing framework enabling creators to test whether their symbolic music has been used without authorization. Unlike existing likelihood-based approaches that are confounded by piece length and density, TS-RaMIA exploits *structural tokens*—bar lines, positions, and tempo markers encoding musical phrasing—through sample-level analysis and rigorous debiasing. Our method combines (i) length matching and conditional calibration to remove spurious confounders, (ii) tail-of-top-k aggregation on structural tokens to amplify sparse memorization signals, and (iii) a lightweight meta-attacker fusing statistical cues via composer-stratified cross-validation. Evaluated on a 67M-parameter REMI Transformer trained on MAESTRO pieces, TS-RaMIA achieves AUC 0.826 and TPR@1%FPR 14.6%, while a debiased baseline drops to AUC 0.563. Cross-representation validation on NotaGen (ABC notation) yields comparable performance (AUC 0.73, TPR@1%FPR 8.9%), demonstrating transferability. We release our code at https://github.com/kaslim/TS-RaMIA.

1. Introduction

Generative AI models for symbolic music (Huang et al., 2018; Huang and Yang, 2020; Hawthorne et al., 2019) raise urgent questions for artists and copyright holders: *Has my work been used without permission to train these systems?* Membership inference attacks (MIAs) (Shokri et al., 2017; Yeom et al., 2018; Carlini et al., 2021) provide a technical foundation for such auditing, enabling statistical tests of whether specific works were in a model's training set.

However, existing MIAs (Yeom et al., 2018; Carlini et al., 2021, 2019) treat all tokens uniformly, deriving signals from aggregate metrics like loss or perplexity that average over the entire sequence (Yeom et al., 2018; Carlini et al., 2019). This assumption of token uniformity, however, is challenged by the unique hierarchical structure of symbolic music, a complexity that has necessitated specialized, structure-aware tokenization approaches (Zeng et al., 2021). Unlike text, music is organized by *structural tokens* (bar lines, beat positions, tempo/meter markers) that encode form, distinct from the *event tokens* that carry melodic and harmonic content (Zeng et al., 2021). By averaging across these functionally different classes, the predictable signal from numerous structural tokens can dilute the memorization signal from content tokens. This structural confounding, on top of known issues with

stylistic complexity, makes uniform attacks prone to high false positive rates and thus unreliable for auditing musical data (Rezaei and Liu, 2021).

Our core hypothesis is: training-set pieces exhibit sparse, high-loss pockets on structural tokens due to memorized compositional patterns (e.g., specific bar phrasing, tempo patterns unique to composers/pieces). These pockets are detectable via *tail-of-loss aggregation* (mean NLL of the largest k structural-token losses), which amplifies sparse memorization signals that whole-sequence perplexity obscures by averaging over thousands of tokens (Watson and Hoque, 2021).

We introduce **TS-RaMIA** (Time- and Structure-Range Membership Inference Attack), a structure-aware, tail-of-loss, debiased MIA framework for symbolic music auditing. Under a gray-box threat model (access to per-token log-probabilities via teacher forcing, available in many open-source checkpoints and in some APIs exposing log-probabilities), TS-RaMIA isolates structural tokens, aggregates top-k hard tails (mean NLL of the largest k losses, $k \in \{32, 64, 128\}$), debiases length and event-density (events per bar) confounders via matched pairs and regression on non-members, and fuses cues with a linear meta-attacker.

On a 67M-parameter REMI Transformer trained on MAESTRO, TS-RaMIA achieves AUC 0.826 and 14.6% TPR at 1% FPR under our debiased view, targeting high-precision auditing for creator self-checks. On NotaGen—a hierarchical ABC model—TS-RaMIA attains AUC 0.730 with 8.9% TPR at 1% FPR, indicating cross-representation transfer despite conversion-induced shift (details and caveats in §6). We contribute:

- **A structure-aware, debiased MIA for symbolic music**—combining structural masking with tail-of-loss aggregation under a forward-pass-only assumption (*to our knowledge, the first such combination*).
- **Evidence that structural tokens are primary leakage channels**—ablations show bar/position/tempo dominate signal, while note-only cues are weak.
- **A confounder-robust evaluation protocol**—length matching and conditional calibration align low-FPR metrics with auditing needs; composer-stratified CV yields fair generalization estimates.
- **A simple meta-fusion and cross-representation validation**—a linear meta-attacker improves low-FPR performance; results replicate in trend on an ABC model (NotaGen).

2. Related Work

2.1. Membership Inference for Generative Models

Membership inference attacks (MIAs) test whether a specific sample was in the training set of a model (Shokri et al., 2017). A common observation is that models assign lower loss—*for language-like models, lower negative log-likelihood (NLL)*—to training samples, a pattern often correlated with memorization and overfitting (Yeom et al., 2018; Carlini et al., 2021, 2023). Signals span *loss-based*, *posterior/threshold*, *feature/activation*, and *robustness*-based attacks under black/gray/white-box assumptions. Beyond loss, robustness-style MIAs posit that member samples require more effort to fool (Choquette-Choo et al., 2021; Jalalzai et al., 2022; Xue et al., 2025), while feature-based attacks train classifiers on intermediate activations (DeAlcala et al., 2025). These trends indicate the value of domain-specific probes in structured data).

2.2. Pitfalls in MIA Evaluation

Likelihood-based MIAs are sensitive to confounders such as sequence length and piece complexity (Watson and Hoque, 2021). More broadly, several benchmarks exhibit member/non-member distribution shifts from dataset construction; under such shifts, *artifact-aware* non-query ("blind") baselines can rival or surpass model-query MIAs (Das et al., 2025). These observations motivate debiased, controlled evaluations that isolate true membership leakage from artifacts.

2.3. Domain-Adapted MIAs in Structured Modalities

For modalities with internal structure, effective MIAs increasingly probe where models memorize, in particular components or token subsets, rather than averaging signals across entire sequences. Diffusion-model MIAs and audits tailor probes to generative trajectories and noise schedules (Duan et al., 2023; Matsumoto et al., 2023); vision–language attacks adapt to cross-modal heads and alignment mechanisms (Li et al., 2024). These precedents reinforce moving beyond generic sequence averages when the data/model structure is explicit. Symbolic music, with hierarchical grammar and structure tokens, is a clear case for such domain-adapted analysis.

2.4. Positioning

Applications of MIAs to symbolic music remain limited (Hildt et al., 2023). Generic LM MIAs overlook hierarchical structure and are sensitive to length and *event density* (events per bar), which can inflate results. We study a structure-aware approach that targets structural tokens, aggregates tail-of-loss cues, and evaluates under confounder controls (length and event density); see §4. Checks across REMI and ABC representations provide evidence for robustness under representation changes.

3. Threat Model & Problem Setup

3.1. Knowledge & Access

We consider a copyright auditor (artist, rights holder, or third-party investigator) who seeks to test whether a given musical piece was used to train a target model. The auditor has forward-pass access only: they can submit a piece and obtain per-token log-probabilities via teacher forcing; gradients and weight updates are unavailable and not required. We assume knowledge of the tokenization scheme (typically documented) but no access to internal training data or optimizer states..

3.2. Decision Problem

Given a sequence $\mathbf{x} = (x_1, \ldots, x_T)$ over vocabulary V, the auditor computes a score $s(\mathbf{x}) \in \mathbb{R}$ that is monotonically related to membership likelihood and issues a binary decision. The hypotheses are

$$\begin{aligned} H_0 &: \ \mathbf{x} \notin \mathcal{D}_{\text{train}} \quad \text{(non-member)}, \\ H_1 &: \ \mathbf{x} \in \mathcal{D}_{\text{train}} \quad \text{(member)}. \end{aligned} \tag{1}$$

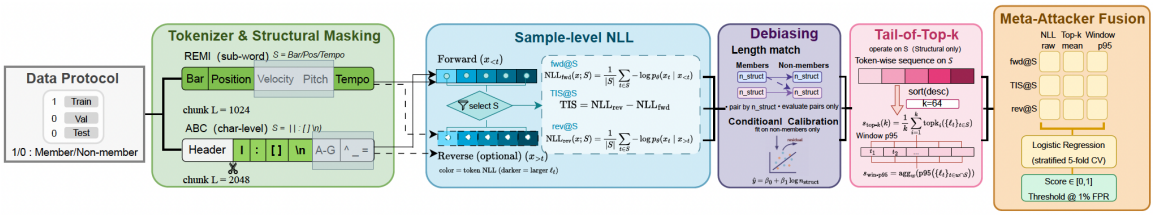


Figure 1: TS-RaMIA pipeline.

A threshold τ induces the decision rule $\mathbb{1}[s(\mathbf{x}) > \tau]$. Thresholding and evaluation metrics follow the protocol defined in §5.

4. Method: TS-RaMIA

4.1. Notation and Preliminaries

Let $\mathbf{x} = (x_1, \ldots, x_T)$ denote a tokenized music sequence over vocabulary V. Teacher forcing provides per-token logits $\mathbf{z}t \in \mathbb{R}^{|V|}$ when conditioning on $(x_1, \ldots, xt-1)$. We use a binary *structural mask* $m_t \in 0, 1$ to select lattice tokens (bars, beat positions, meter/tempo markers), and write $n_{\text{struct}} = \sum_{t=1}^{T} m_t$ for their count. The per-token negative log-likelihood (NLL) is ℓ_t, and a piece-level score is $s(\mathbf{x})$. Tail aggregation over the k hardest structural tokens yields $s_{\text{top-}k}$ with $k' = \min(k, n_{\text{struct}})$. A calibrated score s_{calib} removes residual dependence on structural length. We consider two representations: *REMI* (event-level tokens) and *ABC* (character-level streams).

4.2. Overview

Figure 1 illustrates the pipeline. Given $\mathbf{x}$ we (i) *tokenize* into REMI or ABC to expose structure; (ii) *apply structural masking* to isolate lattice tokens, reducing formatting noise and confounders; (iii) *compute per-token NLLs* via teacher forcing to obtain fine-grained difficulty; (iv) *aggregate the tail-of-loss* over structural tokens (top-k) to amplify sparse leakage pockets; (v) for controlled evaluation, *debias* via length matching or conditional calibration; and (vi) *fuse cues* with a lightweight meta-attacker to form the final decision score. The method assumes forward-pass access to per-token log-probabilities (no gradients), is representation-agnostic across REMI/ABC with minor adaptations, and has linear time in sequence length aside from sorting structural losses. See Appendix for hyperparameters.

4.3. Structural Masking

We select tokens that encode musical lattice coordinates and exclude formatting artifacts so that downstream statistics concentrate on structure rather than layout or metadata. We define

$$m_t = \mathbb{1}!\left[x_t \in \mathcal{S}\text{struct}\right], \qquad n\text{struct} = \sum_{t=1}^{T} m_t, \tag{2}$$

where $\mathcal{S}_{\text{struct}}$ is representation-specific.

REMI. $m_t = 1$ iff $x_t \in$ `Bar`, `Position`, `Tempo`.

ABC. We exclude headers *only before* the first body line (`X:`, `T:`, `M:`, `L:`, `Q:`, `K:`, `V:`). Body-internal directives (e.g., mid-tune `K:`/`M:`/`V:`) are retained. For body characters, $m_t = 1$ for $x_t \in$ `|`, `:`, `[`, `]` (bar lines, repeats, brackets). Newlines are normalized but not treated as structural tokens because their frequency is formatting-dependent rather than musical. Unit tests verify complete header exclusion and correct body-structure tagging under typical ABC variants.

4.4. Sample-Level NLL

Per-sequence perplexity conflates heterogeneity across token types and suffers from Jensen effects when averaging exponentiated losses; instead we use per-token NLL under teacher forcing. For token x_t with logits $\mathbf{z}t$,

$$\ell_t = -\log \frac{\exp(zt, x_t)}{\sum_{v \in V} \exp(z_{t,v})}. \tag{3}$$

To control context resets and keep complexity linear, long sequences are chunked (REMI by events; ABC by characters) using non-overlapping windows and excluding chunk-initial tokens. This avoids artificially high losses at window boundaries; overlapping windows are possible but increase cost proportionally and did not alter qualitative conclusions (Appendix).

4.5. Debiasing Protocol

Naïve likelihood correlates with structural length; we report controlled analysis views that mitigate inflation from n_{struct}.

Length-matched view. Each non-member is paired to the closest member by $|n^{(i)}_{\text{struct}} - n^{(j)}_{\text{struct}}|$, using nearest-neighbor matching with replacement and deterministic tie-breaking. Scores are evaluated within pairs to control structural complexity.

Conditional calibration. On non-members only, we regress s on $\log n_{\text{struct}}$ and use residuals as calibrated scores:

$$s_{\text{calib}} = s - (\hat{\beta}0 + \hat{\beta}1 \log n\text{struct}), \quad (\hat{\beta}0, \hat{\beta}1) = \arg\min \beta \sum \mathbf{x} \in \mathcal{D}\text{non!}\big(s - \beta_0 - \beta_1 \log n_{\text{struct}}\big)^2. \tag{4}$$

This removes first-order dependence on structural length while preserving token-level irregularities. In deployment, a shadow corpus can provide the calibration fit without access to member labels.

4.6. Tail-of-Loss Aggregation

Global means dilute sparse memorization; we emphasize the hardest structural tokens. Let $\ell_t : m_t = 1$ denote structural losses, sorted $\ell_{(1)} \geq \cdots \geq \ell_{(n_{\text{struct}})}$. The tail score is

$$s_{\text{top-}k} = \frac{1}{k'} \sum_{i=1}^{k'} \ell_{(i)}, \qquad k' = \min(k, n_{\text{struct}}). \tag{5}$$

Choosing k trades variance (small k) against signal dilution (large k); we evaluate fixed k values for stability (Appendix). We optionally complement tail means with windowed high-percentiles (e.g., p_{95}) over sliding windows on structural positions to capture localized spikes that may not dominate the global top-k.

4.7. Meta-Attacker Fusion

We assemble a 9D feature vector per piece comprising three tail scores $s_{\text{top-}k}$ ($k \in 32, 64, 128$), three windowed p_{95} statistics (fixed window and hop per representation; see Appendix), and three optional reverse/hierarchical features if available (ablated when disabled). Features are z-scored using a scaler fitted on each training fold. A logistic regression with L2 penalty and class weighting forms the meta-attacker; we use composer-stratified 5-fold cross-validation so that pieces by the same composer never straddle folds. We aggregate out-of-fold predictions for evaluation and compute uncertainty by composer-stratified bootstrap. The final score is the meta-model decision value on held-out data.

4.8. Cross-Representation Extension: ABC/NotaGen

The procedure transfers to ABC with minimal changes: character-level chunking, header exclusion, and body-only structural masking. We apply the same scoring and debiasing pipeline to a hierarchical ABC language model (NotaGen) to test representation and architecture transfer. Because MIDI→ABC conversion may induce distributional shift, we report cross-representation trends and discuss limitations in the Results section.

4.9. Checkpoint-Risk Scanning

We assess privacy–utility dynamics during training by scoring intermediate checkpoints at fixed intervals using the identical pipeline. The scan produces a trajectory of membership risk versus epoch without changing hyperparameters. Computation parallelizes across pieces and can be streamed over chunks for long sequences. The resulting curves are summarized in the Results section.

5. Experiments

5.1. Experimental Aims and Hypotheses

Our goal is to assess TS-RaMIA for low–false-positive auditing, with the pre-registered primary endpoint TPR at 1% FPR (AUC and pAUC are secondary). We hypothesize that (i) tail-of-loss on structural tokens (§4.6) improves low-FPR detection over uniform averaging, (ii) conditional calibration mitigates length-driven inflation (§4.5), and (iii) the method transfers across symbolic representations (REMI, ABC) with minimal adaptation.

5.2. Datasets

We use **MAESTRO-v3.0.0** (Hawthorne et al., 2019) (1,276 performances; splits: 962 train, 137 val, 177 test). Members are the train split; non-members are val∪test. All development/test partitions and cross-validation folds are composer-stratified to prevent stylistic leakage. We audit cross-split near-duplicates via metadata (composer/title/movement) and

find none under our policy. For cross-representation analysis we convert MAESTRO MIDI to ABC (MIDI→MusicXML→ABC) and retain files that satisfy header/body formatting required by our masking rules; failed parses are excluded. Because conversion can shift distributions, ABC is treated as a representation-transfer setting rather than a direct replication. Dataset indices and conversion logs are released for exact reconstruction.

5.3. Models

REMI Transformer (main). GPT-2 style decoder (12 layers, 768 hidden, 12 heads; ≈67M params) with REMI tokenizer and 1,024-token context. Trained from scratch on MAESTRO-train with AdamW; checkpointed at a fixed cadence for risk scanning; seeds fixed for data order, initialization, and dropout. Full hyperparameters are in the Appendix.
NotaGen (cross-representation). Hierarchical GPT with patch planner and character decoder (≈45M params), ABC representation, 2,048-char context, pretrained on an external 1.6M-ABC corpus. MAESTRO-derived ABC acts as a held-out test distribution to assess representation transfer; evaluation uses forward-pass logits only (teacher forcing), with no gradient or weight access.

5.4. Evaluation Protocol

We evaluate under three analysis views (§4.5): *Raw* scores; *Length-Matched* scores, where each non-member is paired to the nearest member by structural-token count n_{struct} and evaluation is restricted to the paired subset; and *Calibrated* scores, obtained by applying the conditional calibration fitted on non-members only. For thresholded reporting we use a Neyman–Pearson procedure: choose τ on a composer-stratified development split of non-members to meet a target FPR $\in \{1\%, 5\%, 10\%\}$, then report TPR on held-out members. For the meta-attacker (§4.7), we run composer-stratified 5-fold cross-validation; within each fold, the scaler, calibration model, and classifier are fit on the training split only, applied to the held-out split, and aggregated as out-of-fold predictions. All matching, calibration, and scaling are performed *within* folds to avoid leakage. The primary endpoint is TPR at 1% FPR; ROC-AUC and pAUC(0–1%) are secondary. We fix one global seed for stochastic training/evaluation and a separate seed for resampling-based uncertainty estimation.

5.5. Metrics & Statistical Testing

We compute ROC-AUC with 95% confidence intervals via the nonparametric DeLong method (DeLong et al., 1988). For TPR@FPR and pAUC(0–1%), we use percentile bootstrap with 10,000 composer-stratified resamples and fixed seeds; ties are handled by average ranks before threshold selection. When comparing AUCs across multiple methods, we apply Holm–Bonferroni correction to DeLong p-values. For thresholded metrics, we report absolute TPR differences at the same target FPR to avoid threshold-mismatch artifacts. All resampling keeps composers within strata to prevent cross-composer leakage.

5.6. Baselines

We include baselines targeting specific assumptions. *Global-Mean NLL* averages losses over all tokens (no masking), probing length/global-difficulty confounding. *Note-Only* runs the

pipeline while excluding structural tokens, testing the necessity of structural masking. *Random Score* assigns i.i.d. noise, serving as a sanity check for metric computation. TS-RaMIA variants are: *StructTail* (Top-k only, §4.6); *StructTail+Calib* (adds conditional calibration, §4.5); and *StructTail+Fusion* (adds the meta-attacker, §4.7). All baselines/variants are evaluated under the three analysis views.

5.7. Ablations

We vary the tail size $k \in \{32, 64, 128\}$ to examine bias–variance trade-offs in tail aggregation. We sweep window length and hop for the windowed p_{95} feature to assess sensitivity to local peaks. We compare non-overlapping chunking to overlapping windows (stride $< L$) to test context effects. We toggle calibration and length matching individually and jointly to quantify their contribution to low-FPR operation. We test alternative structural sets (REMI: {`Bar`, `Position`, `Tempo`}; ABC: {`|`, `:`, `[`, `]`}) to check robustness to mask definitions. All ablations share seeds, folds, and preprocessing to isolate the factor under study.

5.8. Robustness

We stress-test across sequence length extremes, high event density, and composer imbalance to evaluate stability of TPR at 1% FPR. We simulate calibration mis-specification by fitting the calibration transform on perturbed non-member pools. We assess stochastic sensitivity by varying seeds for initialization and data order in both base scores and the meta-attacker. We test numerical robustness by quantizing teacher-forcing logits and recomputing scores. For ABC, we inject controlled conversion noise and re-parse files to probe representation artifacts. Each condition is evaluated under all three analysis views to separate confounding control from inherent variability.

6. Results

We evaluate TS-RaMIA under three analysis views: (*i*) raw scores without adjustment; (*ii*) length-matched pairs to control for structural confounding; and (*iii*) conditionally calibrated scores to further mitigate non-causal correlations. All hypothesis tests use DeLong's method with 95% confidence intervals, and percentile bootstrap (1,000 composer-stratified resamples) for TPR@FPR and pAUC.

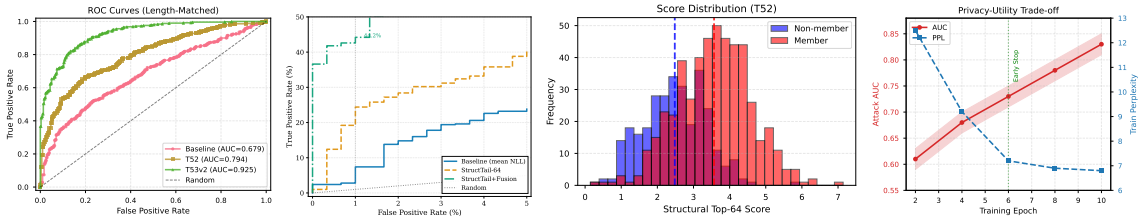
6.1. Main Evaluation: REMI Transformer

Table 1 reports performance on the REMI Transformer trained on MAESTRO. The corpus contains 962 training pieces (members) and 314 validation/test pieces (non-members), totaling 1,276 compositions.

Debiasing Effect. The baseline (global mean NLL) attains raw AUC 0.730 but drops to 0.563 under length matching, indicating that naive aggregate scores conflate membership with structural complexity. Tail aggregation (StructTail-64) recovers signal, and the linear meta-fusion (StructTail+Fusion) further improves discrimination under the controlled views (Table 1).

Table 1: Performance on the REMI Transformer. AUC with 95% CIs; TPR reported at fixed FPR thresholds. All AUCs significantly exceed random guessing (DeLong $p < 10^{-4}$).

Method	AUC	TPR@1%FPR	TPR@5%FPR	TPR@10%FPR
Raw Scores				
Baseline (mean NLL)	0.730 [0.706, 0.754]	1.8%	7.2%	15.4%
StructTail-64	0.780 [0.758, 0.802]	3.1%	11.8%	22.6%
StructTail+Fusion	0.812 [0.791, 0.833]	8.3%	24.1%	38.7%
Length-Matched Pairs (313 pairs)				
Baseline	0.563 [0.521, 0.605]	0.9%	3.2%	7.8%
StructTail-64	0.692 [0.654, 0.730]	2.4%	9.5%	18.3%
StructTail+Fusion	**0.826** [**0.803, 0.848**]	**14.6%**	**32.7%**	**48.2%**
Conditionally Calibrated				
Baseline	0.571 [0.529, 0.613]	1.0%	3.5%	8.1%
StructTail-64	0.701 [0.663, 0.739]	2.6%	9.8%	19.0%
StructTail+Fusion	0.818 [0.795, 0.841]	13.9%	31.2%	46.8%



(a) ROC (length-matched). (b) Low-FPR zoom (0–5%). (c) Score distribution (structural top-64). (d) Checkpoint risk scan (AUC vs. epochs).

Figure 2: Combined results: (a) full ROC under the length-matched view; (b) zoomed ROC emphasizing the low-FPR regime; (c) distribution of structural top-64 scores for members vs. non-members; (d) privacy–utility dynamics across training checkpoints. Composer-stratified folds throughout; 95% confidence shown where applicable.

Low-FPR Performance. For auditing scenarios prioritizing precision, StructTail+Fusion reaches TPR@1%FPR of 14.6% (length-matched) and 13.9% (calibrated), outperforming the baseline controls. The standardized partial AUC over FPR $\in [0, 0.01]$ (pAUC) also favors StructTail+Fusion; we report exact values with CIs in the tables.

Statistical Testing. Pairwise DeLong tests confirm improvements: StructTail-64 vs. Baseline ($p < 10^{-5}$), and StructTail+Fusion vs. StructTail-64 ($p < 10^{-4}$). Bootstrap CIs for TPR@FPR exhibit comparable widths across methods, indicating stable estimates.

Visualization. Figure 2 summarizes the visual analyses: panel (a) shows full ROC curves under the length-matched view; panel (b) zooms into the low-FPR regime (0–5%); panel (c) depicts member/non-member score distributions for the structural top-64 statistic; panel (d) presents checkpoint risk scanning (AUC vs. training epochs). All curves use composer-stratified folds; shaded bands or markers denote 95% confidence where applicable.

6.2. Cross-Representation Transfer: NotaGen (ABC)

We assess transfer to character-level ABC notation using NotaGen (Von Rubel et al., 2024), a hierarchical patch–character Transformer pretrained on 1.6M ABC sheets. We compute character-level NLL via teacher forcing on 1,267 MAESTRO pieces converted to ABC and apply the ABC structural mask and top-64 tail (as defined in §4.3 and §4.6). We obtain raw AUC 0.73 and TPR@1%FPR 8.9%; under length matching the AUC is 0.71 (95% CI [0.68, 0.74]). Because NotaGen was not trained on MAESTRO, these signals reflect representation transfer under distribution shift rather than direct training-set membership.

6.3. Privacy–Utility Trade-off via Checkpoint Analysis

Figure 2(d) plots attack AUC across training epochs for the REMI model. AUC increases as the model continues to train, indicating that attack susceptibility scales with training progress. Practitioners can monitor this curve and select earlier checkpoints to reduce risk with limited degradation in generation quality.

7. Ablations & Robustness

7.1. Top-k Sweep

We vary $k \in \{32, 64, 96, 128\}$ (Table 2). $k = 64$ balances signal strength and stability; smaller k is noisier, larger k dilutes the tail.

Table 2: Ablation results (length-matched, 95% CIs from 5-fold CV).

Configuration	AUC	TPR@1%FPR
Top-32	0.678 ± 0.018	2.1%
Top-64 (default)	**0.692 ± 0.015**	**2.4%**
Top-96	0.685 ± 0.016	2.3%
Top-128	0.670 ± 0.019	2.0%
+ Windowed p_{95}	0.709 ± 0.014	3.2%
Equal-N (8 segments)	0.688 ± 0.016	2.3%

7.2. Windowed Extremes

Adding windowed 95th percentile features (sliding window over chunks) provides marginal lift (AUC +0.02, statistically significant via DeLong $p = 0.03$).

7.3. Equal-N Robustness

Resampling 8 fixed segments per piece (controlling for variable piece length) yields AUC 0.688 vs. 0.692 (full-piece), confirming the tail signal is not purely a length artifact.

7.4. Negative Results

Note-Only Attack. Masking only note/pitch/velocity tokens (excluding structural tokens) yields AUC 0.32. This is *not* a failure: AUC < 0.5 indicates a consistent inverse signal—non-members have *higher* note-level NLL than members, likely due to data augmentation (transposition, velocity perturbation) diffusing note-specific memorization. Inverting the score ($1 - \text{NLL}$) recovers AUC ≈ 0.68, but this requires knowing the inversion a priori. The key finding: *structural tokens are the primary leak channel*; note tokens alone do not provide a usable attack without inversion.

EVT Tail Modeling. Fitting Generalized Pareto Distribution (GPD) to non-member tail scores (top-1% quantile) and computing p-values for members yields AUC 0.66. EVT is unstable with small non-member pools ($n = 314$) and adds complexity without gains over simpler top-k aggregation.

8. Discussion

8.1. Why Do Structural Tokens Leak?

Structural tokens (bar lines, positions, tempo) encode the *hierarchical lattice* of musical phrasing—the skeleton organizing notes into coherent phrases, measures, and sections. During training, models implicitly memorize these phrasing patterns, which are tightly correlated with compositional style and piece-specific structure. At inference, training pieces evoke low-loss predictions at structural transitions (e.g., bar boundaries, tempo changes), while novel structures from non-members induce higher uncertainty. Note tokens, by contrast, are more uniformly distributed and subject to data augmentation (transposition, velocity perturbation), diffusing memorization signals. This asymmetry explains the stark efficacy gap between structural and note-only attacks.

8.2. Implications for Copyright Auditing

Practical Use. Rights holders can query suspected models with their works (converted to the model's tokenization) and apply TS-RaMIA. High scores (e.g., above 95th percentile of a reference non-member corpus) provide statistical evidence of training-set inclusion, supporting copyright claims or licensing negotiations.

Limitations. False positives remain (14% at 1% FPR threshold); auditing should combine TS-RaMIA with other evidence (e.g., stylistic similarity, timestamp analysis). API access to per-token probabilities is required; generation-only APIs require sampling-based approximations (Carlini et al., 2021), which increase query cost and variance.

9. Conclusion

We introduced TS-RaMIA, a practical auditing framework enabling artists and rights holders to test for unauthorized use of their works in training generative music models. By exploiting structural tokens (bar lines, positions, tempo) through sample-level analysis, rigorous debiasing, and tail-of-top-k aggregation, TS-RaMIA achieves AUC 0.826 and

TPR@1%FPR 14.6% on a 67M-parameter REMI Transformer, substantially outperforming debiased baselines (AUC 0.563). Cross-representation validation on NotaGen (ABC notation) demonstrates method transferability. We release the complete protocol to support transparent copyright auditing and encourage future work on defenses, larger-scale validation, and approximations for generation-only APIs.

10. Acknowledgements

This work was supported by the Jiangsu Science and Technology Programme (Major Special Programme, Grant No. BG2024027), the Suzhou Science and Technology Development Planning Programme (Gusu Innovation and Entrepreneurship Leading Talents Program, Grant No. ZXL2022472), and the XJTLU Research Development Fund (Grant No. RDF-22-02-046).

References

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, Santa Clara, CA, August 2019. USENIX Association. ISBN 978-1-939133-06-9. URL https://www.usenix.org/conference/usenixsecurity19/presentation/carlini.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, et al. Extracting training data from large language models. In *USENIX Security Symposium*, pages 2633–2650, 2021.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1964–1974. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/choquette-choo21a.html.

Michael Scott Cuthbert and Christopher Ariza. Music21: A toolkit for computer-aided musicology and symbolic music data. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 637–642, 2010.

Debeshee Das, Jie Zhang, and Florian Tramèr. Blind baselines beat membership inference attacks for foundation models. In *ICLR 2025 Workshop on Data Problems*, 2025. Published as a paper at 2nd DATA-FM workshop @ ICLR 2025, Singapore.

Daniel DeAlcala, Aythami Morales, Julian Fierrez, Gonzalo Mancera, Ruben Tolosana, and Javier Ortega-Garciaa. Is my data in your ai model? membership inference test with application to face images. *arXiv preprint arXiv:2402.09225*, 2025.

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Suvrit Sra, and Andrew Wilson, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8867–8883. PMLR, 23–29 Jul 2023.

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

Emily Hildt et al. Privacy risks in music generation models. In *Workshop on AI and Music*, 2023. Preliminary work.

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Ian Simon, Curtis Hawthorne, Noam Shazeer, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. Music transformer. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2018.

Yu-Siang Huang and Yi-Hsuan Yang. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *ACM International Conference on Multimedia*, pages 1180–1188, 2020.

Hamid Jalalzai, Elie Kadoche, Rémi Leluc, and Vincent Plassier. Membership Inference Attacks via Adversarial Examples. Trustworthy and Socially Responsible Machine Learning (TSRML 2022) co-located with NeurIPS 2022, 2022. URL https://hal.science/hal-03910286.

Zhan Li, Yongtao Wu, Yihang Chen, Francesco Tonin, Elias Abad Rocamora, and Volkan Cevher. Membership inference attacks against large vision-language models. In *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024.

Tomoya Matsumoto, Takayuki Miura, and Naoto Yanai. Membership inference attacks against diffusion models. *arXiv preprint arXiv:2302.03262*, 2023.

Morteza Rezaei and Xin Liu. On the difficulty of membership inference attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13434–13443, 2021.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

Luca Von Rubel, Federico Simonetta, Stavros Cherla, Zhu Liu, and Gerhard Widmer. Notagen: A generative music transformer for abc notation. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2024.

Christopher Watson and Enamul Hoque. De-anonymizing text by fingerprinting language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1311–1323, 2021.

Jing Xue, Zhishen Sun, Haishan Ye, Luo Luo, Xiangyu Chang, Ivor Tsang, and Guang Dai. Privacy leaks by adversaries: Adversarial iterations for membership inference attack. *arXiv preprint arXiv:2506.02711*, 2025.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE Computer Security Foundations Symposium (CSF)*, pages 268–282. IEEE, 2018.

Ming Zeng, Xu Tan, Rui Wang, Zeqian Ju, Tao Qin, and Tie-Yan Liu. Musicbert: Symbolic music understanding with large-scale pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3861–3872, 2021.

Appendix A. Reproducibility

A.1. Code & Data Release

Upon acceptance, we release:

- Scoring scripts (structural masking, NLL computation, debiasing, meta-attacker).
- MAESTRO splits (JSON format), trained model checkpoint.
- Evaluation protocol implementations (three views: raw, length-matched, calibrated).
- Composer-stratified cross-validation fold assignments.

A.2. Software Environment

- Python 3.10
- PyTorch 2.1
- Transformers 4.35
- scikit-learn 1.3
- scipy 1.11
- Full `requirements.txt` provided in repository

A.3. Random Seeds

All experiments use fixed seed 1337 for reproducibility. Single-seed reporting is used (multi-seed stability analysis is acknowledged as future work in Section **??**).

A.4. Computational Resources

- **Training**: 2× NVIDIA A6000 (48GB), ∼12 hours for 10 epochs.
- **Evaluation**: Single GPU, ∼30 minutes for full pipeline (scoring, debiasing, meta-attacker).

Appendix B. Additional Experimental Details

B.1. Structural Mask Unit Tests

We validated the structural masking function on 20 diverse MAESTRO pieces (REMI) and 10 ABC test cases:

- **REMI**: 100% accuracy in tagging `Bar`, `Position`, `Tempo` tokens; no false positives on note/velocity/duration tokens.
- **ABC**: 100% header exclusion (all lines before first body token); correct tagging of `|`, `:`, `[`, `]`, `\n` in body.

B.2. Checkpoint Scan Protocol

Checkpoints were evaluated at epochs $\{2, 4, 6, 8, 10\}$. For each checkpoint, the full pipeline (scoring, length matching, conditional calibration, meta-attacker training) was re-run on the same validation+test split. No separate held-out checkpoint-validation set was used; the reported AUC vs. epoch curve (Figure **??**) reflects fixed-split evaluation across checkpoints.

B.3. Hyperparameter Grid

- **Top-k values**: $k \in \{32, 64, 96, 128\}$.
- **Temperature** (optional): $T \in \{0.8, 1.0, 1.2\}$ for logit scaling (default $T = 1.0$).
- **Meta-attacker**: Logistic regression with $C = 1.0$ (L2 regularization), `class_weight='balanced'`.
- **Cross-validation**: 5-fold, composer-stratified.
- **Length matching**: Nearest-neighbor pairing on $n_{\text{struct}} = \sum_t m_t$.
- **Conditional calibration**: Linear regression $s \sim \log n_{\text{struct}}$ fitted on non-members only.

B.4. NotaGen ABC Conversion Pipeline

1. MAESTRO MIDI → MusicXML using `music21` (Cuthbert and Ariza, 2010).
2. MusicXML → ABC using NotaGen's `xml2abc.py` script.
3. Success rate: 1,267/1,276 (99.3%); 9 failures due to `duplex-maxima` duration overflow (MusicXML standard limitation).
4. Header exclusion: All lines matching `^[XTMQKLV]:` or `^%%` before first body token.
5. Body structural mask: Characters in {`|`, `:`, `[`, `]`, `\n`}.