How Far Should Tokenization Go?
Predictive Effectiveness and Relational Losslessness

Anonymous authors
Paper under double-blind review

Abstract

GPT-style models have achieved remarkable success with finite vocabularies of reusable tokens, making the token interface a central component of modern sequence modeling. Symbolic music appears naturally compatible with this paradigm: it consists of discrete note events and recurring structures such as chords, motifs, and phrases. However, when tokenization moves beyond language, the interface must be specified for each domain. Existing work offers many effective designs, but no unified criterion for deciding what tokenization should represent and how far it should go.

Using predictive codelength as a common criterion, we formulate the *Effectiveness–Losslessness Framework* to define where tokenization should begin and where it should end. The *Fact–Token Boundary* marks where observation-determined structure should enter the token interface, through operations such as coordinate construction. Within this interface, the resulting carrier may be reversibly recoded, for example through BPE, without changing the represented facts. The *Token–State Boundary* marks where tokenization should stop: relations that depend on context should remain for model-state computation rather than being fixed in advance by the tokenizer.

We validate the framework through controlled multi-seed symbolic-music experiments, with an independent-corpus replication of the temporal intervention. Making musical time explicit consistently reduces predictive code and also improves pitch and duration prediction, while tonal-frame canonicalization and pitch factorization provide further gains. Fixed circle-of-fifths pitch coordinates instead increase predictive code, suggesting that imposing a fixed pitch relation before context can burden prediction. Reversible BPE substantially shortens the carrier but increases predictive codelength in every seed, showing that carrier compaction alone does not guarantee predictive gain.

Together, these results show that tokenization quality is determined neither by recoverability nor by sequence length, but by where predictive computation is placed. The framework therefore recasts tokenization as a model-relative boundary: static representation exposes reusable regularities, while contextual state resolves what remains context-dependent.

1 Introduction

1.1 Tokenization in GPT-Style Models

GPT-style models have achieved remarkable success by combining large-scale sequence modeling with finite vocabularies of reusable tokens (Vaswani et al., 2017; Brown et al., 2020). The token interface maps text into a discrete sequence of units that can be reused across contexts, making tokenization a central component of the paradigm. Because language tokenizers are mature and widely standardized, their design can appear secondary to the model architecture.

1.2 Symbolic Music and Reusable Structure

Symbolic music appears naturally compatible with the same paradigm. It consists of discrete Note and REST events and contains recurring structures such as chords, motifs, phrases, and sections.

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.

Existing systems instantiate this idea through event streams, compound words, learned segments, bar/position timing, and absolute or relative pitch–time interfaces (Huang et al., 2018; Huang & Yang, 2020; Zeng et al., 2021; Fradet et al., 2023a; Guo et al., 2023; Inaba et al., 2024). Yet musical structure also spans simultaneity, meter, transposition, and long-range recurrence, which are not captured by a single serialization axis.

1.3 The Missing Criterion for Tokenization

Existing work offers many effective tokenization designs, but no unified criterion for deciding what tokenization should represent and how far it should go.

Comparisons between complete tokenizers cannot supply such a criterion. Interfaces that preserve the same information may still impose different predictive burdens on a finite model (Xu et al., 2020; Rajaraman et al., 2024; Jafari Fesharaki et al., 2026). Moreover, a change in predictive codelength may arise from coordinate construction, reversible carrier coding, decoder or compute changes, or a static relational prior. Predictive compression depends on whether a representation exposes recurring regularities as reusable conditional structure; recoverability and shorter sequences do not guarantee this.

We therefore use predictive codelength as a common criterion for measuring representation operations and determining where static tokenization should begin and end.

Contributions.

1. We make operation-level tokenization decisions measurable under a common predictive-codelength criterion. By separating coordinate construction, reversible carrier recoding, and fixed relational commitment, we define operation-specific quantities that identify how each intervention changes the predictive burden of a bounded learner.
2. Building on this criterion, we formulate the *Effectiveness–Losslessness Framework* with two complementary principles. Predictive Effectiveness defines the *Fact–Token Boundary*, where observation-determined structure should enter the token interface; Relational Losslessness defines the *Token–State Boundary*, where tokenization should stop before context-dependent relations are fixed in advance.
3. We validate the resulting predictions through controlled symbolic-music experiments. Observation-determined coordinate operations produce predictive gains, fixed circle-of-fifths pitch coordinates increase predictive burden, and reversible BPE substantially shortens the carrier without producing predictive gain. An independent corpus replicates the temporal-coordinate direction.

2 The Effectiveness–Losslessness Framework

2.1 Predictive Codelength as a Common Criterion

Predictive compression arises when a representation makes recurring regularities appear as stable, reusable conditional structure to the specified learner; emitting fewer symbols alone does not guarantee this. We therefore compare tokenization operations by the predictive data codelength achievable under a fixed source, learner family, and resource budget.

We separate the declared fact surface, coordinate construction, carrier coding, and contextual computation:

$$x_{\mathrm{raw}} \xrightarrow{A} x \xrightarrow{R} z_{1:N} \xrightarrow{C} u_{1:T} \xrightarrow{F_\theta} h_{1:T}. \tag{1}$$

A declares the modeled fact surface x, which may abstract the raw domain. R constructs coordinate carriers $z_{1:N}$, C serializes or groups them as $u_{1:T}$, and F_θ forms contextual states $h_t = F_\theta(u_{<t})$ for $q_\theta(u_t \mid h_t)$.

For source distribution P, predictor family $\mathcal{Q}$, and resource budget $\mathcal{B}$, we define

$$\mathcal{L}_{P,\mathcal{Q},\mathcal{B}}(R, C) = \frac{\inf_{q \in \mathcal{Q}_\mathcal{B}} \mathbb{E}_{x\sim P}\left[-\sum_{t=1}^{T} \log_2 q(u_t \mid u_{<t})\right] + \mathbb{E}_{x\sim P}[\ell_{\mathrm{side}}(x)]}{\mathbb{E}_{x\sim P}[N_{\mathrm{fact}}(x)]}, \quad u = C(R(x)). \tag{2}$$

The resource budget $\mathcal{B}$ fixes the learner and evaluation conditions under which representations are compared, including model capacity, context, training exposure, and selection policy. We measure predictive data codelength rather than full MDL; model description length is outside this ledger. Accordingly, all codelength differences are relative to the specified source, predictor family, and budget (Xu et al., 2020).

Matched intervention quantities. We distinguish three representation operations under the same predictive criterion. Coordinate construction changes the model-facing coordinates, carrier coding changes their reversible serialization, and relational commitment imposes a fixed relation before contextual computation.

Using $\mathcal{L}$ as shorthand for fixed $(P, \mathcal{Q}, \mathcal{B})$, we define

$$\begin{aligned} G_{\text{coord}}(R_0 \to R_1) &= \mathcal{L}(R_0, C) - \mathcal{L}(R_1, C), \\ G_{\text{carrier}}(C_0 \to C_1) &= \mathcal{L}(R, C_0) - \mathcal{L}(R, C_1), \\ E_{\text{rel}}(g) &= \mathcal{L}(R_g, C) - \mathcal{L}(R_{\text{open}}, C). \end{aligned} \tag{3}$$

Positive G_{coord} and G_{carrier} indicate predictive gains. Positive E_{rel} indicates excess predictive code from the tested fixed relation. Each quantity is interpreted only under a matched intervention in which the corresponding operation changes and the remaining comparison conditions are held fixed.

2.2 Predictive Effectiveness: The Fact–Token Boundary

The first decision is what structure should enter the token interface. The *Predictive Effectiveness Principle* admits an observation-determined operation when exposing it reduces predictive codelength for the specified learner.

Information preservation does not imply computational equivalence. Two representations may preserve the same declared facts while imposing different approximation burdens on a bounded learner. Invertibility is therefore insufficient to establish coordinate usefulness.

The Fact–Token Boundary therefore marks where tokenization begins. A tested candidate is predictively effective when it yields positive G_{coord} under a matched resource contract.

Two general coordinate operations are central in this work: decoupling and denesting.

Decoupling. For a declared fact $x = (a, b)$, a recoverable factor-wise interface has

$$R_{\text{dec}}(a, b) = \big(r_A(a), r_B(b)\big), \qquad \text{Rec}_{\text{dec}}(R_{\text{dec}}(a, b)) = (a, b). \tag{4}$$

The factors become separately addressable, but no independence is imposed: $P(A, B) \neq P(A)P(B)$ may still hold. The operation removes only the representational constraint that the two factors must be accessed through a coupled coordinate, while leaving their statistical dependence available to the contextual predictor. Its effectiveness requires matched positive gain $G_{\text{coord}}(R_{\text{coupled}} \to R_{\text{dec}}) > 0$; recoverability alone does not determine the sign.

Denesting. For an observation-determined transformation ϕ, a recoverable exposed interface has

$$R_\phi(x) = \big(\phi(x), s_\phi(x)\big), \qquad \text{Rec}_\phi(R_\phi(x)) = x. \tag{5}$$

Here $s_\phi(x)$ is the information required for inversion and may be empty when ϕ is bijective; coded side information is charged through ℓ_{side} in equation (2). Moving ϕ into the interface is predictively effective only when $G_{\text{coord}}(R_0 \to R_\phi) > 0$ under the matched source, learner, and budget. Determinism and reversibility alone do not imply this gain.

A particularly important case of denesting is *canonicalization*, in which ϕ aligns a known transformation orbit. This structured form removes nuisance variation while preserving the information required to return to the original coordinates, without assigning a contextual semantic relation.

These operations are candidates for the static interface because they are determined by declared facts; their predictive value is measured, not assumed.

Deterministic relations may also enter the interface. Exact signed displacement between preserved coordinates is observation-determined. The Fact–Token Boundary is therefore not between facts and relations, but between structure determined by observation and organization that requires context.

2.3 Reversible Carrier Coding

Once a coordinate interface is admitted, its contents may be serialized or grouped in different reversible ways without changing the represented coordinates. Common tokenization operations such as BPE act at this carrier level by merging or regrouping symbols into a different token sequence.

Carrier compaction is not predictive compression:

$$|C_1(z)| < |C_0(z)| \;\not\Rightarrow\; \mathcal{L}(R, C_1) < \mathcal{L}(R, C_0).$$

A reversible recoding may shorten the emitted sequence while making prediction no easier, or even harder, for the specified learner. Its value is therefore measured by predictive carrier gain rather than sequence length alone.

2.4 Relational Losslessness: The Token–State Boundary

The second decision is where tokenization should stop. The *Relational Losslessness Principle* concerns organization whose useful form cannot be determined from the preserved items alone.

For two items x_i and x_j, suppose their useful relation depends on surrounding context:

$$\rho_{ij}^{\mathcal{C}} = f(x_i, x_j, \mathcal{C}), \qquad \exists\, \mathcal{C}_1, \mathcal{C}_2 : \rho_{ij}^{\mathcal{C}_1} \neq \rho_{ij}^{\mathcal{C}_2}. \tag{6}$$

The same preserved pair can therefore support different useful organizations under different contexts.

A static interface makes a *relational commitment* when it instead privileges a context-free relation or topology

$$g_{ij} = g(x_i, x_j) \tag{7}$$

before the relevant context is available. A *relation-open* interface preserves the identities or coordinates needed to infer such relations without fixing this tokenizer-side geometry in advance.

Crucially, relational losslessness is not invertibility. A committed representation may preserve every declared fact,

$$R_g^{-1}(R_g(x)) = x,$$

yet still increase predictive burden:

$$R_g^{-1}(R_g(x)) = x \;\not\Rightarrow\; E_{\mathrm{rel}}(g) \leq 0. \tag{8}$$

Under the specified source, predictor, and budget, $E_{\mathrm{rel}}(g) > 0$ is *relational coding loss*: the fixed commitment requires more predictive code than its matched relation-open reference.

The *Token–State Boundary* therefore marks where context-dependent organization should remain unresolved by the static interface and instead be determined through contextual state. This does not imply that every fixed relation is harmful; a commitment is ruled out only when it produces relational excess under the matched predictive criterion.

2.5 The Unified Tokenization Boundary

Together, the two principles define tokenization as a model-relative boundary over predictive computation. Predictive Effectiveness moves observation-determined structure into the static interface when doing so reduces predictive burden, while Relational Losslessness stops this process before context-dependent organization is prematurely fixed.

A token interface is therefore a carrier of admitted observable structure, not a commitment to every higher-order organization that can be formed from it. Reversible carrier coding acts within this interface and is evaluated separately through G_{carrier}. Figure 1 summarizes the resulting division of predictive computation across the static interface, carrier code, and contextual state.

> *Tokenization should expose reusable structure determined by observation and stop before precomputing organization that context must determine.*

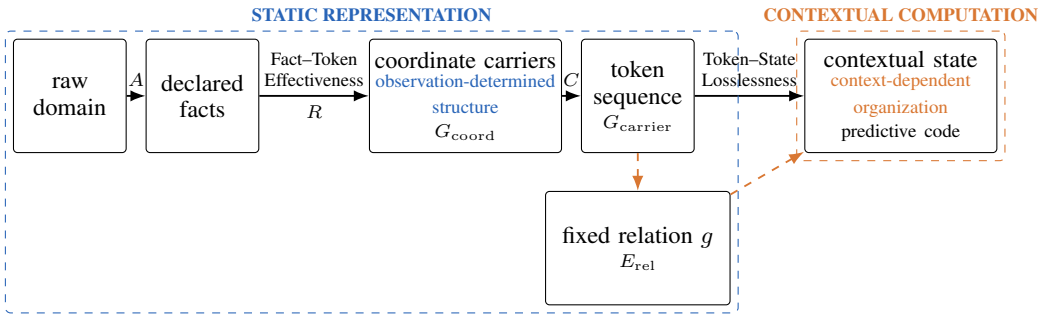

Figure 1: The Effectiveness–Losslessness Framework. Predictive Effectiveness governs what enters the token interface, Relational Losslessness governs where it stops, and reversible carrier coding is evaluated separately.

3 Symbolic Music as a Controlled Testbed

Symbolic music makes the framework concrete because score-level facts are explicit while many useful organizations remain context-dependent. We use this separation to derive coordinate, relational, and carrier predictions before specifying their matched experimental realizations.

3.1 Declared Fact Surface

Under our score abstraction, each observed event is

$$n_i = (t_i, o_i, p_i, d_i), \tag{9}$$

where t_i is event type, o_i onset, p_i pitch, and d_i duration. For a REST, onset and duration delimit the silent interval. These variables define the declared fact surface.

Expressive performance and low-level MIDI attributes lie outside this abstraction. All modeled non-drum notes form one unlabeled event stream. Same-onset events remain distinct, and serialization order assigns neither melodic priority nor voice identity.

3.2 Predictive Effectiveness in Symbolic Music

Musical time. Sequence position reflects serialization rather than musical time. We instead expose each onset through a directed multiscale metric coordinate

$$\tau_i = \phi(o_i) = \big(\text{bar progress}, \text{beat/bar/multibar phases}\big), \tag{10}$$

and derive exact signed temporal displacements between events from their observed onsets. These operations expose score-determined temporal structure without assigning any contextual musical organization.

Predictive Effectiveness predicts positive coordinate gain when this exposure makes reusable temporal regularities easier for the bounded predictor to access. Because time participates in the joint event distribution, the gain need not be confined to Time prediction; it may also appear in pitch, duration, or event type.

Pitch factorization. Absolute pitch can be written bijectively as pitch class and register,

$$p_i \longleftrightarrow (c_i, r_i), \qquad c_i = p_i \bmod 12, \qquad r_i = \lfloor p_i/12 \rfloor . \tag{11}$$

This decouples two reusable factors without assuming that they are statistically independent. Predictive Effectiveness therefore asks whether making them separately addressable lowers predictive codelength; invertibility alone does not determine the answer.

Tonal canonicalization. Transposition defines a known transformation orbit over absolute pitch. Choosing a deterministic reference frame gives

$$\widetilde{p}_i = p_i - s, \qquad p_i = \widetilde{p}_i + s, \tag{12}$$

with the inverse shift retained outside the predictor. This is a structured denesting operation: it removes a nuisance degree of freedom without supplying tonic, mode, scale degree, or harmonic function as semantic labels.

The predicted signature is positive coordinate gain after charging the information required for inversion.

3.3 Relational Losslessness in Symbolic Music

Pitch provides a controlled test of the Token–State Boundary because pitch identity can be preserved while its model-facing geometry is changed. A relation-open interface exposes pitch class c_i as a categorical identity without prescribing a fixed distance between pitch classes.

A relation-committed interface instead maps pitch class to a fixed circular coordinate

$$\gamma_\kappa(c_i) = \left[\sin \frac{2\pi\kappa(c_i)}{12}, \cos \frac{2\pi\kappa(c_i)}{12}\right], \tag{13}$$

where $\kappa(c) = c$ gives chromatic order and $\kappa(c) = 7c \bmod 12$ gives circle-of-fifths order. This mapping remains injective, but its geometry privileges proximity under the chosen topology before musical context is available.

Relational Losslessness predicts $E_{\text{rel}} > 0$ when such a fixed geometry precommits pitch organization that would be better resolved from context.

3.4 Reversible Carrier Coding in Symbolic Music

The same admitted musical coordinates can be serialized through different reversible carriers. Event symbols such as bar, position, pitch, and duration may be emitted directly or regrouped by a learned merge scheme such as BPE without changing the recoverable coordinate grammar.

The carrier hypothesis is deliberately weaker than a compression claim: shortening the emitted sequence need not reduce predictive codelength. The appropriate empirical signature is therefore G_{carrier}, evaluated with the represented coordinates held fixed.

4 Related Work

Predictive tokenization under bounded models. Tokenization and representation mappings have been formalized independently of any particular vocabulary (Gastaldi et al., 2025). Predictive usable information and bounded-model analyses show that information-preserving representations can impose different predictive burdens on computationally bounded learners (Xu et al., 2020; Rajaraman et al., 2024; Jafari Fesharaki et al., 2026; Bechler-Speicher et al., 2026). Related work also shows that token count or sequence compaction alone does not determine predictive quality (Zouhar et al., 2023; Schmidt et al., 2024). These results motivate predictive codelength as our accounting criterion; our focus is the operation-level distinction between coordinate construction, carrier recoding, and relational commitment.

Musical coordinates and relational structure. Symbolic-music models expose bar/position timing, temporal and pitch relativity, and combinations of absolute and relative musical attributes (Huang et al., 2018; Huang & Yang, 2020; Fradet et al., 2023b; Guo et al., 2023; Inaba et al., 2024; Guo & Dixon, 2025). These works provide precedents for placing domain structure into the model-facing interface. We distinguish observation-determined coordinate exposure from context-free relational commitment and evaluate the two through matched predictive interventions.

Carrier coding and sequence compaction. Compound words, BPE, and learned merging change how musical information is carried by the token sequence (Zeng et al., 2021; Fradet et al., 2023a; Qu et al., 2025; Geerlings & Meroño-Peñuela, 2020; Pasquier et al., 2025). Such recoding can alter sequence length, vocabulary, and the effective resources available per declared fact. We treat reversible carrier coding as a separate operation: represented coordinates are held fixed, and shorter serialization is not assumed to imply lower predictive codelength.

Compression and contextual abstraction. Hierarchical music models organize computation across contextual scales, while segment-based tokenizers move higher-level abstractions into the token interface (Yu et al., 2022; Huang et al., 2025). DeCo similarly separates token compression from semantic abstraction in multimodal modeling (Yao et al., 2024). These works motivate the broader allocation question addressed here: which structure should be made static and reusable, and which organization should remain for contextual state.

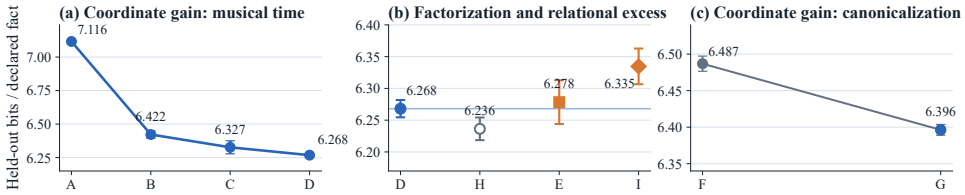

Figure 2: Pop1K7 clean-test codelength per declared musical fact. Error bars show sample SD over three preregistered seeds. Panel (a) reports the staged A–D musical-time family. Panel (b) contains separate matched comparisons: D/H estimates pitch-factorization gain; D/E and H/I estimate relational excess for chromatic and fifths commitments, respectively. Panel (c) reports the matched F/G canonicalization family. G includes a conservative four-bit side code per window.

5 Experiments and Discussion

5.1 Matched Experimental Design

We instantiate the three operation classes in equation (3) through matched Pop1K7 comparisons:

Quantity	Comparison	Deliberate change
G_{coord}	A–D	musical-time coordinates
G_{coord}	D–H	pitch factorization
G_{coord}	F–G	tonal canonicalization
E_{rel}	D–E	fixed chromatic geometry
E_{rel}	H–I	fixed fifths geometry
G_{carrier}	J–K	reversible BPE recoding

Within A–D, A uses generic sequence position, B exposes absolute musical position, C replaces it with multiscale musical time, and D additionally exposes exact temporal displacement. The remaining comparisons change only the operation named above within their matched family.

Only matched within-family deltas receive operation-level interpretation; cross-family scores are descriptive. All Pop1K7 families use a 16-layer, width-64 causal backbone and a 2,000 target-equivalent-epoch maximum budget; full parameter and compute contracts appear in appendices A and I.

Pop1K7 (Huang & Yang, 2020) is split by source-song lineage before windowing. All primary comparisons use three preregistered seeds, validation-only checkpoint selection, and one sealed clean-test pass. Codelength is normalized by declared Note/REST facts and includes EOS and required side information; G is charged four inverse-shift bits per independent window.

5.2 Predictive Coordinate Gain

Musical time. Codelength decreases at every stage of A–D, with gains of 0.694, 0.096, and 0.059 bits/fact for A→B, B→C, and C→D. Overall,

$$G_{\text{coord}}(A \to D) = 0.84809 \text{ bits/fact},$$

an 11.92% reduction, with all three seeds favoring D (figure 2a).

Only 0.287 bits/fact of this gain comes from Time prediction; roughly two thirds of the gain occurs in the other event factors, principally Pitch and Duration. Thus exposing musical time reorganizes the joint conditional distribution rather than merely making timestamps easier to predict. The complete decomposition appears in appendix D.

The A/D direction also replicates in all three ComMU seeds (Lee et al., 2022), with a mean gain of 0.111 bits/fact. Because the runs were still improving at the budget limit, this establishes directional rather than converged replication (appendix J).

Pitch-coordinate operations. After charging the inverse shift, tonal canonicalization yields $G_{\text{coord}}(F \to G) = 0.09055$ bits/fact, with all three seeds improving. The operation removes a

transposition degree of freedom without supplying tonic, mode, scale degree, or harmonic function. Pitch factorization yields a smaller gain of 0.03171 bits/fact; two seeds improve and one is approximately tied (figure 2b,c).

5.3 Relational Excess Codelength

The primary Token–State test compares relation-open H with the fixed circle-of-fifths geometry I. Although both preserve pitch class and register, the fixed topology incurs

$$E_{\text{rel}}(\text{fifths}) = \mathcal{L}(I) - \mathcal{L}(H) = 0.09831 \text{ bits/fact},$$

with all three seeds regressing. Pitch contributes 0.08428 bits/fact of this excess, locating most of the penalty in the representation that was changed.

The chromatic D/E comparison is much smaller ($E_{\text{rel}} = 0.01032$ bits/fact) and has mixed seed directions. It therefore shows no reliable gain from the tested fixed chromatic geometry, rather than the seed-consistent loss observed for fifths.

A separate fixed-target probe provides complementary state-side evidence: removing either ordered context increases code by about 0.27 bits/event, while shuffling the same content increases it by 0.669 bits/event (appendix G). This supports context-sensitive use of preserved evidence but is not a matched Token–State intervention.

5.4 Reversible Carrier Coding

The J/K comparison holds the recoverable coordinate grammar fixed while K applies reversible BPE to J's serialized stream. BPE reduces emitted targets from 3.563 to 1.538 per fact (56.8%), yet every seed requires more predictive code:

$$G_{\text{carrier}}(J \to K) = -0.81061 \text{ bits/fact}.$$

Thus substantial carrier compaction does not imply predictive compression under the tested learner and budget. Full carrier-induced vocabulary, parameter, and compute changes are reported in appendix I.

5.5 Scope and Limitations

All effects are predictive-data-codelength results relative to the tested source, learner, and resource budget, and mechanistic attribution is restricted to matched families. The results do not imply that all fixed relations or BPE schemes are harmful; ComMU provides directional rather than converged replication, and the context probe is supporting evidence only. We do not establish perceptual quality, interpretable hidden-state variables, or asymptotic scaling.

6 Conclusion

How far should tokenization go? Our central answer is that tokenization is a model-relative allocation of predictive computation. The *Effectiveness–Losslessness Framework* makes this allocation operational: Predictive Effectiveness admits observation-determined structure when exposing it reduces predictive codelength, while Relational Losslessness stops static representation before context-dependent organization is fixed in advance. Reversible carrier coding is evaluated separately. Tokenization quality is therefore determined neither by recoverability nor by sequence length alone, but by where predictive computation is placed.

Symbolic music provides controlled evidence for each part of this distinction. Musical-time coordinates reduce predictive code by 11.92%, with roughly two thirds of the gain appearing outside Time prediction; tonal canonicalization remains beneficial after its inverse is charged, and pitch factorization yields a smaller gain. In contrast, an injective fixed circle-of-fifths geometry produces seed-consistent relational excess, while reversible BPE shortens the carrier by 56.8% yet increases predictive codelength by 0.811 bits/fact. An independent corpus reproduces the temporal-coordinate direction. Together, these results support an operation-level view of tokenization: expose reusable structure, preserve contextual relational freedom, and do not equate carrier compaction with predictive compression.

These results suggest a revealing view of representation: tokens compress what can be standardized in advance, while contextual states learn how to compress what must remain context-dependent. From this perspective, self-attention is powerful not merely because it accesses context, but because it can construct the predictive representation itself as context changes.

AI Use Statement

Generative AI systems, including OpenAI Codex/ChatGPT and Google Gemini, were used as research-assistance tools for conceptual brainstorming and critical discussion, code drafting and refactoring, experiment orchestration, literature discovery, figure and table preparation, and editorial revision. AI outputs were not treated as scientific evidence. The human authors retained final control over the research questions, experimental protocols, data inclusion, interpretations, and claims; verified the cited sources, implementations, and reported results; and take full responsibility for the submission. Formal results were produced through the frozen deterministic protocols and auditable receipts reported in the paper rather than from unverified model-generated values.

Reproducibility Statement

Reproducibility is addressed in section 5 and appendices A, G, I and J. They specify data lineage, declared facts, numerator and denominator accounting, side-information codes, matched interventions, active parameters, budgets, seeds, and checkpoint rules. Protocol, source-tree, split-manifest, and checkpoint hashes bind each result to its frozen implementation and data lineage. Test data were unavailable to training and checkpoint selection; the sealed test was materialized once after checkpoint lock. The J/K technical continuation used only hash-verified caches and did not reopen the test manifest or row file.

Ethics and Data Statement

The primary representation and carrier-coding experiments use Pop1K7 (Huang & Yang, 2020); ComMU (Lee et al., 2022) provides an independent temporal-coordinate replication. Both are used for non-commercial research with attribution, and we make no claim of ownership over source compositions. No human-subject study or listener data collection was conducted. Public data-derived artifacts will follow the applicable upstream terms.

References

Maya Bechler-Speicher, Gilad Yehudai, Gil Harari, Clayton Sanford, Amir Globerson, and Joan Bruna. Lost in tokenization: Fundamental trade-offs in graph tokenization for transformers. *arXiv preprint arXiv:2605.22471*, 2026. doi: 10.48550/arXiv.2605.22471.

Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 2020.

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. Association for Computational Linguistics, 2023a. doi: 10.18653/v1/2023.emnlp-main.123. URL https://aclanthology.org/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 *Proceedings of the 24th International Society for Music Information Retrieval Conference*, 2023b. URL https://archives.ismir.net/ismir2023/paper/000009.pdf.

Juan Luis Gastaldi, John Terilla, Luca Malagutti, Brian DuSell, Tim Vieira, and Ryan Cotterell. The foundations of tokenization: Statistical and computational concerns. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=B5iOSxM2I0.

Cariña Geerlings and Albert Meroño-Peñuela. Interacting with GPT-2 to generate controlled and believable musical sequences in ABC notation. In *Proceedings of the 1st Workshop on NLP for Music and Audio*, pp. 49–53. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.nlp4musa-1.10/.

Zixun Guo and Simon Dixon. Moonbeam: A MIDI foundation model using both absolute and relative music attributes. *arXiv preprint arXiv:2505.15559*, 2025. doi: 10.48550/arXiv.2505.15559.

Zixun Guo, Jaeyong Kang, and Dorien Herremans. A domain-knowledge-inspired music embedding space and a novel attention mechanism for symbolic music modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5070–5077, 2023. doi: 10.1609/aaai.v37i4.25635.

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, and Douglas Eck. Music transformer: Generating music with long-term structure. *arXiv preprint arXiv:1809.04281*, 2018.

Jingyue Huang, Zachary Novack, Phillip Long, Yupeng Hou, Ke Chen, Taylor Berg-Kirkpatrick, and Julian McAuley. MuseTok: Symbolic music tokenization for generation and semantic understanding. *arXiv preprint arXiv:2510.16273*, 2025. doi: 10.48550/arXiv.2510.16273.

Yu-Siang Huang and Yi-Hsuan Yang. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

Tatsuro Inaba, Kazuyoshi Yoshii, and Eita Nakamura. On the importance of time and pitch relativity for transformer-based symbolic music generation. In *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2024. doi: 10.1109/APSIPAASC63619.2025.10849230.

Amirmehdi Jafari Fesharaki, Mohammadamin Rami, and Aslan Tchamkerten. Effective context in transformers: An analysis of fragmentation and tokenization. *arXiv preprint arXiv:2605.13485*, 2026. doi: 10.48550/arXiv.2605.13485.

Hyun Lee, Taehyun Kim, Hyolim Kang, Minjoo Ki, Hyeonchan Hwang, Kwanho Park, Sharang Han, and Seon Joo Kim. ComMU: Dataset for combinatorial music generation. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/fff3ba5059aeeb88c324b6ba9b298166-Abstract-Datasets_and_Benchmarks.html.

Philippe Pasquier, Jeff Ens, Nathan Fradet, Paul Triana, Davide Rizzotti, Jean-Baptiste Rolland, and Maryam Safi. MIDI-GPT: A controllable generative model for computer-assisted multitrack music composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1474–1482, 2025. doi: 10.1609/aaai.v39i2.32138.

Xingwei Qu, Yuelin Bai, Yinghao Ma, et al. MuPT: A generative symbolic music pretrained transformer. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=iAK9oHp4Zz.

Nived Rajaraman, Jiantao Jiao, and Kannan Ramchandran. An analysis of tokenization: Transformers under markov data. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1999. URL https://papers.nips.cc/paper_files/paper/2024/hash/724afcaae4ae92a9220a077ffe80088d-Abstract-Conference.html.

Craig W. Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. Tokenization is more than compression. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 678–702. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.40. URL https://aclanthology.org/2024.emnlp-main.40/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/2002.10689.

Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. DeCo: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv preprint arXiv:2405.20985*, 2024. doi: 10.48550/arXiv.2405.20985.

Botao Yu, Peiling Lu, Rui Wang, Wei Hu, Xu Tan, Wei Ye, Shikun Zhang, Tao Qin, and Tie-Yan Liu. Museformer: Transformer with fine- and coarse-grained attention for music generation. *arXiv preprint arXiv:2210.10349*, 2022.

Mingliang Zeng, Xu Tan, Rui Wang, Zeqian Ju, Tao Qin, and Tie-Yan Liu. MusicBERT: Symbolic music understanding with large-scale pre-training. *arXiv preprint arXiv:2106.05630*, 2021.

Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Mrinmaya Sachan, and Ryan Cotterell. Tokenization and the noiseless channel. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5184–5207. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.284. URL https://aclanthology.org/2023.acl-long.284/.

A Frozen Protocol Summary

The A–E, F–G, H–I, and J–K families each have frozen protocol, source-tree, and manifest hashes in the repository receipts. P is the corresponding frozen split, $\mathcal{Q}$ is the shared causal-Transformer family, and $\mathcal{B}$ is the tabled resource contract. Inactive representation paths remain instantiated but frozen; active parameter count is therefore part of the full interface intervention.

Table 1: Frozen source, learner, resource, and accounting contract.

Item	Frozen value
Dataset	Pop1K7 constant-4/4 piano arrangements; source-song split before windowing
Songs / windows	1,378 / 18,460 train; 184 / 2,011 validation; 185 / 1,979 sealed test
Carrier and target	A–I: exact Note/REST stream with a separate termination target and Type→Time→Pitch→Duration factors; J/K: lossless BAR/POSITION/pitch-or-REST/duration stream, with train-only reversible BPE in K
Accounting	numerator: all Note/REST factor bits, EOS overhead, and required per-sample side bits; denominator: declared Note/REST facts only
Backbone	16 Transformer layers, width 64, 4 heads, FFN 128, dropout 0; J/K use the same causal backbone with a serialized-token input/output head
Parameters	815,571 total; effective trainable counts are A 809,475; B 811,079; C 810,499; D/F/G 812,815; E/I 805,455; H 813,519. Vocabulary-dependent J/K totals are 689,349 / 804,675
Context	deterministic complete 8-bar windows, capped at 2,048 note tokens; no random note or bar crop
Seeds	20260819, 20260820, 20260821
Budget	2,000 target-equivalent epochs; 3,650,482,000 full-split exposures; 2,664,430,000 matched-subset exposures. Coordinate/relation families fix decoder and compute policy per fact; J/K fix original-event exposure and the backbone while carrier length and vocabulary change as the intervention
Optimization	AdamW; 3×10^{-4} to 3×10^{-5} cosine schedule with warmup; batch size 128 songs
Selection	minimum native micro validation code; ties favor the earliest exposure. Fact-level rebasing and G's constant side charge do not change checkpoint order
Side information	G shift support $\{-5, \ldots, 6\}$: fixed-length 4-bit code, charged once per 8-bar window in the primary ledger and once per source song in the amortized audit
Admissible deltas	coordinate: A–D, D–H, F–G; relation: E–D, I–H; carrier: J–K only. J–D and K–D are descriptive full-interface comparisons
Test firewall	test data disabled and unloaded during training and checkpoint selection; one manifest and row-file read after checkpoint lock; J/K cache-only technical continuation without a second test access

B Exposure-Matched Optimization Dynamics

The figure is included to distinguish achievable predictive code length from transient optimization effects. Main results therefore use the minimum validation code length under the common maximum

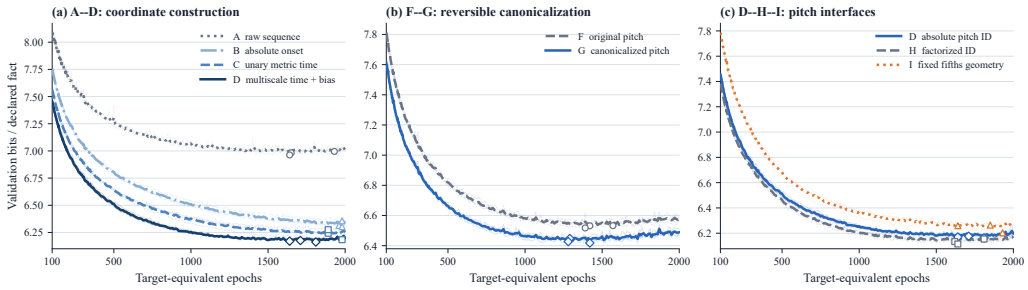

Figure 3: Exposure-matched validation dynamics for the three frozen Pop1K7 representation families. Thin lines show individual preregistered seeds; thick lines show the three-seed mean; open markers identify the independently validation-selected checkpoint for each seed. The plot begins at 100 target-equivalent epochs to remove the initialization transient, while all source observations remain in the released curve table. Every run continues to the same 2,000-epoch maximum budget. A–D and D–H–I use the full split; F–G use their exactly matched high-confidence subset. Curves are rebased to declared facts; G includes the constant four-bit side charge per window. No test value enters training, plotting, or checkpoint selection.

budget, followed by one sealed test evaluation of the frozen checkpoint; they do not compare arms at an arbitrary common epoch.

C Full A–I Arm Table

Table 2: Frozen Pop1K7 representation arms. A–E and H–I use the full original-pitch split; F–G use the exactly matched high-confidence tonal-frame subset. Values are bits per declared fact; G includes the conservative per-window side code. Absolute scores are compared only within matched families.

Family	Arm	Active interface	Validation	Held-out test
Coordinate: time	A	raw pitch identity; generic sequence position	6.98064 ± 0.01633	7.11618 ± 0.00765
Coordinate: time	B	A + learned absolute onset/bar lookup	6.32519 ± 0.02835	6.42223 ± 0.02730
Coordinate: time	C	raw pitch; unary ordered and multi-scale metric time replacing generic/absolute position features	6.23396 ± 0.04671	6.32662 ± 0.04850
Coordinate: time	D	C + pairwise onset, meter, bar, and beat relations	6.16952 ± 0.00681	6.26809 ± 0.01361
Relation: chromatic	E	D time; chromatic-circle phase + absolute register replacing categorical pitch-class identity	6.18494 ± 0.03436	6.27841 ± 0.03446
Coordinate: canonicalization	F	original absolute pitch on the matched music21 high-confidence subset	6.52727 ± 0.00848	6.48686 ± 0.01038
Coordinate: canonicalization	G	deterministic whole-song tonal canonicalization; inverse retained outside model	6.43069 ± 0.01506	6.39631 ± 0.00744
Coordinate: pitch factorization	H	learned categorical absolute pitch class + learned absolute register; D time	6.13428 ± 0.01959	6.23638 ± 0.01785
Relation: fifths	I	fixed circle-of-fifths phase + learned absolute register; D time	6.24231 ± 0.03362	6.33469 ± 0.02811

F and G use the same songs and non-pitch fields from the music21 consensus high-confidence subset. F presents original absolute pitch. G estimates one deterministic whole-song reference frame and canonicalizes pitch according to

$$\widetilde{p}_i = p_i - s, \tag{14}$$

predicts $\widetilde{p}_i$, and restores physical pitch through

$$p_i = \widetilde{p}_i + s. \tag{15}$$

The reference shift is retained for deterministic reconstruction but is not supplied to the Transformer. Dataset key labels, tonic, mode, and harmonic function are unused.

The manifest contains 12 possible shifts, so a fixed-length code costs four bits. The primary result charges this once for each of 1,500 test windows; a source-song decoder can instead reuse one shift across the 132 test songs.

Table 3: Clean-test F/G side-information ledger in bits per declared fact.

Ledger	F code	G code	$G_{\mathrm{coord}}(F \to G)$
Model code only	6.48686	6.36174	0.12512
4 bits / 8-bar window	6.48686	6.39631	0.09055
4 bits / source song	6.48686	6.36478	0.12208

D Full Paired Factor Decomposition

Table 4: Held-out paired changes in bits per declared fact. Negative values favor the left arm. Class is assigned by intervention; Side is nonzero only for G.

Comparison	Class	Total	Type	Time	Pitch	Duration	Side
B−A	G_{coord}	−0.693948	−0.020283	−0.248170	−0.243551	−0.181944	0
C−B	G_{coord}	−0.095607	−0.000855	−0.033450	−0.052890	−0.008412	0
D−C	G_{coord}	−0.058531	−0.000165	−0.005304	−0.043773	−0.009290	0
D−A	G_{coord}	−0.848087	−0.021304	−0.286924	−0.340213	−0.199646	0
E−D	E_{rel}	+0.010321	−0.000059	−0.006559	+0.017385	−0.000445	0
G−F	G_{coord}	−0.090548	+0.000003	+0.002641	−0.137238	+0.009474	+0.034573
H−D	G_{coord}	−0.031711	+0.000028	−0.004057	−0.025560	−0.002121	0
I−H	E_{rel}	+0.098307	−0.000013	+0.010218	+0.084278	+0.003824	0
I−D	descriptive	+0.066596	+0.000015	+0.006161	+0.058718	+0.001703	0

E Split and Test Integrity

Pop1K7 is split by source song before windowing. Exact, transposition-invariant, and rhythm/interval duplicate components remain within one split; all audited cross-split counts are zero. Model selection uses validation only, with test disabled during training.

All 27 frozen checkpoints were evaluated once with no parameter updates. The full test split and matched F/G subset were opened only for this pass.

F Representation Details

All modeled non-drum notes are placed in a single unsegregated event stream. No melody, bass, harmony, accompaniment, instrument-part, or voice partition is supplied to the model. Notes sharing an onset remain separate observable events and are serialized by ascending MIDI pitch. There is no cardinality head, pitch-set decoder, onset-group identity, melody extraction, chord label, or voice label.

For events without pitch, including REST events, a fixed type-specific tie-breaking convention is used. Serialization order is deterministic but does not imply performance order, melodic priority, voice membership, or semantic importance.

Duration remains an exactly recoverable attribute of each Note or REST event rather than an active-note state reconstructed from later transitions. The decoder order is

$$\textsc{Type} \to \textsc{Time} \to \textsc{Pitch} \to \textsc{Duration}. \tag{16}$$

This is a chain-rule factorization of the exact-event distribution rather than a semantic ranking of musical attributes.

C, D, E, H, and I use directed bar progress together with within-beat, beat-within-bar, bar, four-bar, and sixteen-bar unary phases. The directed component disambiguates repeated periodic phases, so the joint musical-time coordinate uniquely determines the supported score onset:

$$\tau_i = \phi(o_i), \qquad \phi^{-1}(\tau_i) = o_i. \tag{17}$$

The unary interface is therefore an invertible coordinate transformation that preserves the supported score onset.

D, E, H, and I additionally use signed onset displacement, within-bar displacement, signed bar distance, and signed beat distance as deterministic attention biases. These pairwise quantities are

computed from preserved musical-time coordinates and do not assign chord, motif, phrase, or voice identity.

D uses raw MIDI-pitch identity. H uses categorical pitch class

$$c_i = p_i \bmod 12 \tag{18}$$

together with categorical absolute register

$$r_i = \left\lfloor \frac{p_i}{12} \right\rfloor . \tag{19}$$

The pair (c_i, r_i) bijectively determines the supported absolute pitch. H is relation-open in the experimental sense: pitch-class identities have learned categorical embeddings, and no context-free pitch-class distance kernel is inserted.

E replaces categorical pitch-class identity with the chromatic phase

$$\gamma_{\mathrm{chr}}(p_i) = \left[\sin\left(\frac{2\pi(p_i \bmod 12)}{12} \right), \cos\left(\frac{2\pi(p_i \bmod 12)}{12} \right) \right] \tag{20}$$

and retains absolute register.

I uses the same phase construction after mapping pitch class to the circle-of-fifths index

$$f_i = \big(7(p_i \bmod 12)\big) \bmod 12, \tag{21}$$

giving

$$\gamma_{\mathrm{fif}}(p_i) = \left[\sin\left(\frac{2\pi f_i}{12} \right), \cos\left(\frac{2\pi f_i}{12} \right) \right] . \tag{22}$$

In E and I, phase inner products privilege chromatic or fifths proximity before contextual attention. The phases remain injective on the twelve supported pitch classes; their intervention class follows from the fixed topology, not from the sign of the result.

F and G use identical songs and non-pitch fields from the music21 high-confidence subset. F predicts original absolute pitch. G applies tokenizer-side whole-song canonicalization,

$$\widetilde{p}_i = p_i - s, \tag{23}$$

predicts canonical pitch, and applies evaluator-side reconstruction,

$$p_i = \widetilde{p}_i + s. \tag{24}$$

The reference shift is retained for deterministic reconstruction but is not supplied to the Transformer. Dataset key labels, tonic, mode, scale degree, and tonal function do not enter G. The primary ledger assigns s a four-bit code for each independent window; the source-song audit amortizes the same code over all windows from one song.

G State-Side Use of Preserved Relations

An earlier relation probe, separate from A–I, tests whether a candidate four-bar target belongs between left and right context. On 6,112 windows from 99 held-out POP909 songs, full-context accuracy was 75.93%, compared with 48.87% for the target-only condition. Replacing left or right context reduced accuracy by 11.42 and 12.37 percentage points, respectively, while replacing both reduced accuracy to 50.87% (figure 4).

Because this probe uses an earlier composition-level representation, it is supporting evidence for contextual dependence rather than part of the primary tokenization result. It is not pooled numerically with the main fixed-target intervention and is not used to support any representation ranking.

A separate fixed-target intervention preserves the target and varies only its available context. Its native evaluation units are retained because this probe uses a different source and task; the values are not pooled with A–I.

Removing either ordered side costs about 0.27 bits/event; shuffling the same amount of content costs 0.6690. The state therefore uses ordered content–time binding rather than context quantity alone.

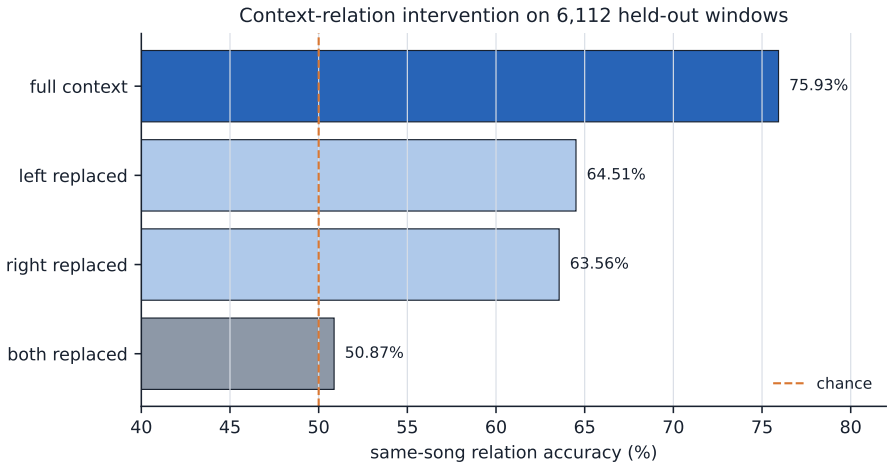

Figure 4: Secondary context-relation probe. Replacing either side degrades compatibility prediction; replacing both approaches the target-only control.

Table 5: Fixed-target context intervention in its native predictive units.

Condition	bits/event	change from Full
FULL	3.5394	—
LEFT	3.8083	+0.2689
RIGHT	3.8117	+0.2723
SHUFFLED	4.2084	+0.6690

H EVIDENCE BOUNDARIES

- The reported data term is a resource-bounded predictive code length, not full MDL or an unrestricted coding optimum. Every comparison is relative to its declared model, optimizer, context, budget, and fact denominator.
- Fact abstraction, reversible carrier coding, and state-side relational commitment are different interventions. We assign a result by what changes, and compare only matched arms within the corresponding family.
- J/K is the admissible carrier intervention: source split, seeds, backbone, original-event exposure, and validation-only selection are fixed while reversible coding changes carrier length and vocabulary. J–D and K–D are descriptive only because event/field autoregression, decoder factorization, Transformer positions per fact, and computation per fact are unmatched.
- H improves over D in two seeds and ties in one. E is mixed; I regresses in all three seeds. These results concern the tested factorization and fixed topologies under slightly different active parameter counts, not every possible pitch geometry.
- F/G uses a matched high-confidence subset and a deterministic whole-song frame. The primary ledger charges four side-information bits per independent window; F/G is not ranked against full-split D and does not test online frame inference.
- Temporal gains establish predictive utility for the tested deterministic time coordinates and relations. The context probes establish sensitivity to ordered content–time binding. Neither identifies a named phrase, cadence, motif, chord, or voice variable in the hidden state.
- Exact recovery is claimed only for the declared Note/REST fields and, where applicable, their separately coded inverse. It does not imply byte-level recovery of raw MIDI, controller data, instrumentation, micro-timing, or acoustic performance.

I POP1K7 FIXED-COORDINATE CARRIER-CODING CONTROL

The frozen serialized evaluator appends one EOS target to every window and includes that target in each arm's loss. Its archived denominator of 244,248 likewise includes 1,979 termination entries.

The paper ledger preserves each arm's complete numerator, including its own EOS cost, and rebases only the denominator to the 242,269 declared Note/REST facts:

$$\mathcal{L}^{\text{paper}} = \mathcal{L}^{\text{archive}} \frac{244{,}248}{242{,}269}. \tag{25}$$

No equality of EOS overhead across J, K, or D is assumed.

Table 6: Pop1K7 fixed-coordinate carrier-coding control. Serialized-target ratios count one EOS target per window; predictive bits include EOS and are normalized by declared Note/REST facts. Values are mean ± sample SD over three preregistered seeds.

Interface	serialized targets/fact	held-out bits/fact
J: uncompressed serialized coordinate stream	3.563	5.96305 ± 0.10235
K: J + train-only reversible BPE	1.538	6.77367 ± 0.05643
$K - J$	-2.025	$+0.81061 \pm 0.15007$

J is an event-preserving REMI-like stream of BAR, exact POSITION, pitch/REST, and duration symbols; it is not a no-coordinate baseline. K applies train-only reversible BPE without changing the recoverable fields or coordinate grammar. The split, seeds, 16-layer backbone, original-event exposure, and validation-only checkpoint rule are matched. Carrier-dependent vocabulary heads bring total parameters to 689,349 for J and 804,675 for K and are part of the tested recoding intervention.

The paired $K - J$ penalties are 0.95195, 0.82678, and 0.65312 bits/fact. Thus $G_{\text{carrier}}(J \to K) = \mathcal{L}(J) - \mathcal{L}(K) = -0.81061$ bits/fact: a 56.8% target reduction produces negative carrier gain under this source, learner family, and budget. Validation selection used no test data; the clean-test manifest and row file were each read once. A technical continuation evaluated hash-verified caches only, with no checkpoint update or test-time BPE fitting.

Descriptive cross-interface reference. J is below D in every seed, whereas K is above D in every seed. These complete system scores are reported transparently but are not admissible coordinate or carrier deltas.

Table 7: Descriptive Pop1K7 full-interface comparisons in the common paper ledger. Differences are left minus right; values are mean ± sample SD.

Comparison	mean difference	seed direction	interpretation
J–D	-0.30504 ± 0.11508	3/3 J below D	descriptive only
K–D	$+0.50558 \pm 0.04366$	3/3 K above D	descriptive only

J and D differ in event- versus field-level autoregression, decoder factorization, Transformer positions and contextual recomputations per original fact, and whether coordinates enter as serialized content or a side interface. J is therefore a strong serialized coordinate interface, not evidence for any single one of those mechanisms. Full-system codelength ranks the tested systems; matched interventions determine why they differ.

J INDEPENDENT-CORPUS A/D REPLICATION

The independent replication uses 9,299 ComMU songs in 4/4, split into 8,652 train, 323 validation, and 324 sealed-test songs. Scores retain their original C-major/A-minor normalization; no second transposition is applied. A and D use the same normalized pitch representation and differ only in their musical-time interface. They otherwise share targets, model envelope, decoder, optimizer, batch construction, context policy, three preregistered seeds, five target-equivalent epochs, and the validation-only checkpoint rule within this corpus. The ComMU budget is not asserted to equal Pop1K7 exposure.

All three paired seeds favor D on validation and sealed test. Test was evaluated once after validation-selected checkpoints were frozen. Exact and transposition-invariant families do not cross splits. Because all six final curves still improve from epoch 4.5 to epoch 5, the result establishes directional replication under a matched budget, not convergence or a ComMU capability limit. On clean test,

Table 8: Frozen ComMU 4/4 A/D replication. Predictive bits include sequence termination and are normalized by declared Note/REST facts. Summary rows report mean ± sample SD over three preregistered seeds; lower is better.

Arm/seed	validation	clean test	test D−A
A: raw sequence	12.56786 ± 0.14342	12.33993 ± 0.14311	—
D: musical-time coordinates	$\mathbf{12.39563 \pm 0.15434}$	$\mathbf{12.22855 \pm 0.18073}$	-0.11139
20260814	—	A: 12.37089; D: 12.34337	-0.02752
20260815	—	A: 12.18388; D: 12.02022	-0.16366
20260816	—	A: 12.46503; D: 12.32205	-0.14298

D primarily improves Type and Time, leaves Duration nearly unchanged, and slightly worsens Pitch. Because the A/D intervention holds pitch encoding fixed, this Pitch difference is an indirect cross-factor effect rather than evidence for or against pitch-coordinate construction. The replication therefore supports the direction of the total temporal-interface effect, not universal improvement of every factor head.