CHORD: LEARNING WATERMARK RECEIVERS FROM UNMARKED PAIRS

Anonymous authors
Paper under double-blind review

ABSTRACT

Watermarks embedded during autoregressive audio generation must support source verification after the generated tokens have been rendered, transformed, and re-encoded. We address this mismatch by learning how the observation channel transforms watermark evidence. For a sender with evaluable likelihood ratios, the observed ratio is the conditional expectation of the source ratio under the unmarked source–observation law. This identity yields a receiver-learning objective and a representation-selection criterion using only paired unmarked data, without waveform reconstruction or key-specific marked training examples. CHORD (CHannel-Observed Response Dictionary) realizes this objective with ordered-candidate sampling, whose rank actions define a small dictionary of local responses shared by all keys and payloads. Learned from unmarked pairs, these responses support composite scoring for joint owner identification and message recovery. On twelve-second music sessions, replacing identity transport with learned responses increases joint owner–payload recovery through two codec compositions from 385 and 379 to 501 and 505 out of 512 sessions, primarily by improving attribution. In speech, unmarked prediction loss selects the receiver, and learned responses yield more gains than regressions under neural codecs, including a held-out operating point. A frozen dictionary transfers to new owner keys without refitting. Together, these results make response learning a practical route to attribution and decoding after audio transformation.

1 INTRODUCTION

Generation-time audio watermarks support provenance only if they remain verifiable after waveform rendering and re-encoding. The recovered token sequence can differ from the original even without additional processing, a problem known as *retokenization mismatch* (Wu et al., 2025a;b). We study verification through this observation process, jointly identifying the generating owner key and recovering its message.

Lossy transformations further complicate verification. Neural codecs can disrupt watermark statistics while retaining useful audio content, and recent assessments report substantial failures of the methods they evaluate (Özer et al., 2025; O'Reilly et al., 2025).

Prior approaches address different parts of the problem. Aligned-IS groups token embeddings and marks clusters; HPT learns token-confusion communities from paired codec passes and biases them during generation (Wu et al., 2025b; Milis et al., 2026). Both seek symbols that remain stable under retokenization. Post-hoc methods learn waveform embedding and recovery, often with transformation augmentation (Chen et al., 2023; San Roman et al., 2024; Liu et al., 2025). We ask instead: *can a receiver learn the channel's effect on a known sender's evidence without training on marked audio?* Figure 1 contrasts these approaches.

The sender can assign a likelihood ratio to an unmarked generation trace from the sampling probabilities of each key–message hypothesis. After the channel, the observed ratio is the conditional average of this source ratio given the observation, under the unmarked joint law. Source traces therefore provide regression targets, and transformed observations provide receiver inputs. Neither waveform reconstruction nor a complete channel model is needed—only its action on verification-relevant functions.

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

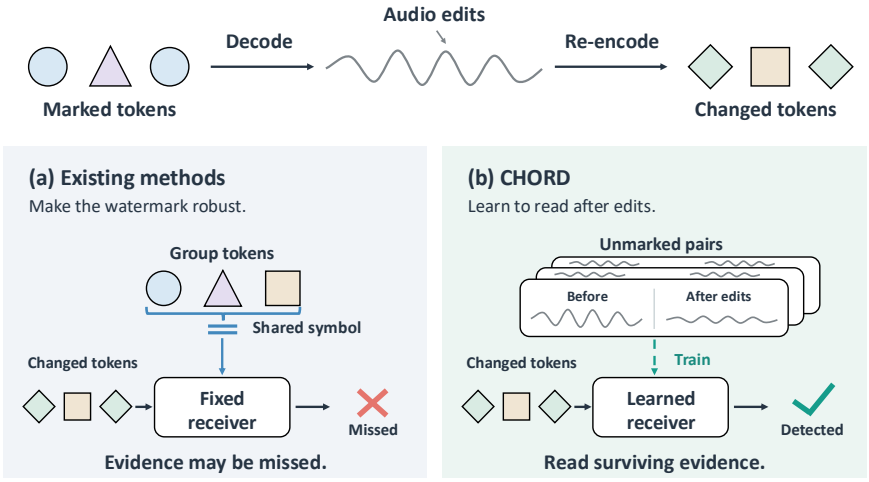

Figure 1: Conceptual comparison of existing methods and CHORD. Edits between decoding and re-encoding can change tokens and disrupt detection. (a) Existing methods improve watermark robustness with a fixed token interpretation. (b) CHORD learns how edits affect watermark evidence from unmarked pairs and stores the result as a response dictionary that its receiver reads.

For distribution-preserving watermarks, averaging over randomized keys recovers the ordinary generation law (Kuditipudi et al., 2024; Hu et al., 2024; Wu et al., 2024). Key-averaged corpora thus supply no binary classification signal, although fixed-key hypotheses remain distinguishable. Learning their responses from unmarked pairs avoids separate marked examples for each owner and message and remains compatible with key-conditioned supervision.

CHORD, a *CHannel-Observed Response Dictionary*, implements this principle. Its ordered-candidate sender retains the token vocabulary and provides closed-form action probabilities and bounded likelihood-ratio targets; the learned post-channel responses of 48 rank actions form a dictionary shared by all keys and payloads. We make three contributions:

- **A receiver-learning formulation.** We identify the shared response operator and why it requires paired data, and derive an unmarked regression objective for estimation and representation selection.
- **A reusable audio realization.** CHORD combines ordered-candidate sampling, a finite response dictionary, and owner–payload scoring, with output-law preservation for the candidate model and response reuse across keys and payloads.
- **Evidence for response learning.** With the sender fixed, learned responses increase twelve-second music recovery through codec compositions from about 75% to 98%, mainly through attribution. We also evaluate unmarked receiver selection, new-key reuse, and transfer across audio channels.

2 Related Work

Generation-time watermarking. Token-level methods make sampling depend on a secret key (Kirchenbauer et al., 2023). Distortion-free sampling, unbiased reweighting, and DiPmark preserve the reference law under their key-randomization conditions while retaining key-aware evidence (Kuditipudi et al., 2024; Hu et al., 2024; Wu et al., 2024). For audio, Aligned-IS uses token-embedding clusters and aligned inverse sampling; HPT builds token-confusion communities from paired codec passes before cluster-level logit biasing (Wu et al., 2025b; Milis et al., 2026). Tokenizer fine-tuning offers another route by improving reverse cycle-consistency (Jovanović et al., 2025). CHORD learns action responses from paired observations instead of merging the sender's symbols.

Waveform watermarking and neural codecs. WavMark, AudioSeal, and XAttnMark support waveform-level embedding and message recovery; AudioSeal also provides localized detection and message-based attribution (Chen et al., 2023; San Roman et al., 2024; Liu et al., 2025). Neural

compression challenges the evaluated methods (Özer et al., 2025; O'Reilly et al., 2025). Latent-Mark directly addresses neural codec compression through waveform optimization targeting latent shifts across surrogate codecs (Chen et al., 2026). CHORD instead learns a shared receiver from unmarked pairs for generation-time sampling and tests transfer across neural codecs.

Likelihood-ratio regression. The conditional-expectation identity is an instance of likelihood-ratio regression, which recovers an observable ratio from tractable latent-variable targets (Brehmer et al., 2020). Conditional-mean operators provide a related view of learning shared transformations from paired samples (Grünewälder et al., 2012). We specialize these ideas to watermark hypotheses, use fixed source targets to compare receiver representations, and implement local response reuse for audio. Composite likelihood describes the local score (Varin et al., 2011); exchangeable-key ranks provide acceptance calibration (Phipson & Smyth, 2010).

3 Learning Watermark Evidence from Unmarked Pairs

3.1 Source Hypotheses and Observations

Let X be a generator-side trace with unmarked law P, and let $P_{m,k}$ be its law under payload $m \in \mathcal{M}$ and key $k \in \mathcal{K}$. The trace includes the context needed to evaluate the sampling probabilities; conditioning information may be fixed, or included with the same distribution under all hypotheses. A key-independent channel $T(y \mid x)$ maps this trace to the receiver's observation Y, incorporating waveform rendering, any subsequent transformation, and re-encoding. The receiver sees the waveform, public session coordinates, and queried keys, but not the source tokens or generation prompt.

Assume $P_{m,k} \ll P$ and that the sender can evaluate its source likelihood ratio and centered *contrast*,

$$w_{m,k}(x) = \frac{P_{m,k}(x)}{P(x)}, \qquad h_{m,k}(x) = w_{m,k}(x) - 1, \tag{1}$$

where the ratios are defined on the support of P. Evaluability is the interface required of the source hypotheses. Section 4.1 defines CHORD's local queries.

Definition 1 (Distribution-preserving watermark)**.** *A keyed sampler is* distribution-preserving *if, for every fixed payload m and conditioning context, $\mathbb{E}_K[P_{m,K}(x)] = P(x)$ for all traces x.*

Here the expectation is over the randomization defining the key; verification therefore compares fixed-key alternatives. The induced observation laws are

$$Q(y) = \sum_x P(x)\, T(y \mid x), \qquad Q_{m,k}(y) = \sum_x P_{m,k}(x)\, T(y \mid x), \tag{2}$$

and our goal is to learn the ratios $Q_{m,k}/Q$ from pairs (X, Y) drawn under ordinary sampling.

3.2 Learn the Channel's Effect on Evidence

For $Q(y) > 0$, inserting $P(x)$ into Equation (2) gives

$$\frac{Q_{m,k}(y)}{Q(y)} = \frac{1}{Q(y)} \sum_x P(x)\, T(y \mid x)\, w_{m,k}(x) = \sum_x \Pr(x \mid y)\, w_{m,k}(x) = \mathbb{E}[w_{m,k}(X) \mid Y = y]\,, \tag{3}$$

so the receiver needs the source ratio averaged over traces compatible with its observation. This is a conditional expectation under the unmarked joint law, not a regression that requires marked examples (Brehmer et al., 2020). Define the response operator

$$\mathcal{R}h(y) = \mathbb{E}[h(X) \mid Y = y] \tag{4}$$

so that $Q_{m,k}/Q = 1 + \mathcal{R}h_{m,k}$. The sender specifies the evidence at generation time; the shared operator specifies its observed form.

Proposition 1 (Learning target)**.** *Suppose the contrasts have finite second moment under P. Then Q together with the values of $\mathcal{R}$ on $\mathcal{H} = \operatorname{span}\{h_{m,k}\}$ determines every ratio $Q_{m,k}/Q$, and two channels with the same P and Q induce the same family of ratios if and only if their response operators agree on $\mathcal{H}$.*

The result follows from linearity (Appendix A.1). It identifies a narrower target than full channel modeling: only the channel's action on the watermark contrasts is needed. CHORD instead learns responses to local action queries for a composite score (Section 4.2).

Pairs are necessary: let X be uniform on $\{-1, +1\}$ and compare $Y = X$ with $Y = -X$. Both channels have identical source and observation marginals, but the admissible contrast $h(x) = \rho x$, $0 < |\rho| < 1$, has response ρy in one and $-\rho y$ in the other. Unpaired samples cannot identify the correct response; pairs retain the needed association.

3.3 Regression and Representation Selection

Fix source-side queries $r_a \geq 0$, indexed by $a \in \mathcal{A}$, with $\mathbb{E}_P r_a = 1$ and finite second moments. Each query defines an alternative $r_a P$. On every unmarked pair, the sender evaluates all targets $r_a(X)$ and the receiver predicts them from Y:

$$L(f) = \mathbb{E} \sum_{a \in \mathcal{A}} \lambda_a \big(r_a(X) - f_a(Y)\big)^2, \qquad \lambda_a > 0, \quad \textstyle\sum_a \lambda_a = 1. \tag{5}$$

The minimizer is $f_a^\star(y) = \mathbb{E}[r_a(X) \mid Y = y]$, the observed ratio for query a. Conditional-mean orthogonality gives

$$L(f) - L(f^\star) = \sum_a \lambda_a \big\| f_a - f_a^\star \big\|^2 \tag{6}$$

with norms in $L^2(Q)$. Held-out pairs estimate the prediction risk $L(f)$ and compare candidate receivers on fixed targets; the optimal risk $L(f^\star)$ need not be known for this comparison.

A receiver may use a representation $Z = g(Y)$ rather than all of Y. Writing $f_{a,Z}^\star = \mathbb{E}[r_a(X) \mid Z]$ yields

$$L(f_Z) - L(f^\star) = \sum_a \lambda_a \Big(\underbrace{\big\| f_a^\star - f_{a,Z}^\star \big\|^2}_{\text{discarded by } g} + \underbrace{\big\| f_{a,Z}^\star - f_{Z,a} \big\|^2}_{\text{not yet estimated}} \Big). \tag{7}$$

The first term is evidence lost by the representation; the second is error in estimating the response available from that representation. With fixed source targets, held-out loss compares both prediction errors without marked validation; composite decisions are evaluated separately in Section 5.2. Appendix A.2 gives the derivation.

The neutral predictor $f_a \equiv 1$ assigns no evidence to any query. Its improvement under a fitted receiver satisfies

$$L(\mathbf{1}) - L(f) = \sum_a \lambda_a \, \chi^2(Q_a \,\|\, Q) \; - \; \sum_a \lambda_a \big\| f_a - f_a^\star \big\|^2, \tag{8}$$

where Q_a is the observation law under $r_a P$. The first term measures surviving query distinguishability; the second is unlearned response error. Since $L(\mathbf{1}) = \sum_a \lambda_a \chi^2(r_a P \,\|\, P)$ is the source-side distinguishability and data processing gives $\chi^2(Q_a \| Q) \leq \chi^2(r_a P \| P)$, the held-out relative reduction $1 - L(f)/L(\mathbf{1})$ estimates a lower bound on the weighted fraction of query evidence that survives the channel and representation (Appendix A.2).

3.4 Attribution and Decoding Use Different Evidence

Let π be a payload prior. An owner's contrast averaged over messages and the residual message-specific contrasts are

$$u_k = \sum_m \pi(m)\, h_{m,k}, \qquad v_{m,k} = h_{m,k} - u_k, \qquad \sum_m \pi(m)\, v_{m,k} = 0. \tag{9}$$

Linearity then gives the owner likelihood ratio and the conditional payload posterior:

$$\ell_k(y) = \sum_m \pi(m) \, \frac{Q_{m,k}(y)}{Q(y)} = 1 + \mathcal{R} u_k(y), \qquad \Pr(m \mid y, k) \;\propto\; \pi(m) \, [1 + \mathcal{R} h_{m,k}(y)] \,. \tag{10}$$

Attribution compares keys through ℓ_k; decoding compares messages within a supplied key. Averaging over payloads cancels the cross term,

$$\sum_m \pi(m) \, \big\| \mathcal{R} h_{m,k} \big\|^2 = \big\| \mathcal{R} u_k \big\|^2 + \sum_m \pi(m) \big\| \mathcal{R} v_{m,k} \big\|^2, \tag{11}$$

so a channel can affect owner evidence and within-owner message variation differently. Correct payload decoding can therefore coexist with weak attribution, which motivates the separate endpoints in Section 5.1.

4 CHORD: A REUSABLE AUDIO WATERMARK RECEIVER

CHORD applies the learning objective to local rank actions and combines their responses into a composite owner–payload score. The generator and native codec stay frozen.

4.1 ORDERED-CANDIDATE SAMPLING

Unlike inverse-transform or Gumbel-based constructions (Kuditipudi et al., 2024; Aaronson, 2022), ordered-candidate sampling supplies a finite alphabet of rank actions with closed-form probabilities. To derive these probabilities, consider a row with context c and draw J candidates independently from the generator's reference conditional p_c. Sort them by a fixed public vocabulary ordering, retaining repeated occurrences. For even J, a keyed phase $U_i \in \{0, \dots, J-1\}$ and payload symbol $z_i \in \{-1, +1\}$ choose the emitted rank

$$u_i = \Big(U_i + \tfrac{J}{2}\,\mathbf{1}\{z_i = +1\}\Big) \bmod J. \tag{12}$$

Proposition 2 (Output-law preservation)**.** *If every marked row draws a fresh phase, uniform on $\{0, \dots, J-1\}$ and independent of the candidates and of the generation history, then the ordered-candidate sender satisfies Definition 1 for every fixed payload.*

A uniform rank selects uniformly among the candidate occurrences, whose expected empirical distribution is p_c. The payload shift only permutes ranks. Repeating this argument at every history proves preservation of the full autoregressive law (Appendix A.3).

Let $F_c(x)$ be the reference cumulative probability in the public order and x^- the predecessor of x. The zero-indexed rank-u candidate is at most x when at least $u+1$ draws are at most x, giving

$$H_{J,u}(v) = \sum_{j=u+1}^{J} \binom{J}{j} v^j (1-v)^{J-j}, \qquad p_c^{J,u}(x) = H_{J,u}\big(F_c(x)\big) - H_{J,u}\big(F_c(x^-)\big). \tag{13}$$

The ratio $p_c^{J,u}(x)/p_c(x)$ lies in $[0, J]$, providing bounded regression targets. Two active codebooks and $J \in \{8, 16\}$ give $2(8+16) = 48$ head–rank actions shared by all keys and payloads.

What the receiver tests. Candidates depend on the unobserved source history, so the receiver cannot recompute them. It tests the phase-conditional hypothesis obtained by averaging over candidate draws: row i follows p_c^{J,u_i}, so the source ratio is a product of local ratios in which (k, m) enters only through u_i. This averaging makes responses shareable: a key and payload select entries of a fixed action dictionary instead of defining new targets. Speech draws candidates from a stream separate from the owner key, so the tested law is the owner-key law marginalized over that stream. Music derives candidates and phases from the same key and nonce through domain-separated HMAC; fixing the key and history fixes the draws, and the averaged law is a surrogate for the fixed-key law. Appendix B.1 specifies both implementations.

4.2 LOCAL RESPONSE LEARNING

A key–payload pair defines a *program* $a_i(k, m)$, one action per generation row. For the action probabilities defined above, the single-action ratio at row i is

$$r_{i,a}(x) = \frac{p_i^a(x_i \mid x_{<i})}{p_i(x_i \mid x_{<i})}, \qquad \mathbb{E}\big[r_{i,a} \mid x_{<i}\big] = 1, \tag{14}$$

and defines the alternative that changes one conditional while leaving the other reference kernels unchanged. This is a local query of Section 3.3: its response is the conditional mean given O_i. A shared dictionary pools examples across rows for reuse across programs.

Music: transition compilation. Map source and observed tokens to finite states using the same public-order coarsening. Estimate a source marginal $\hat{p}$ and transition $\widehat{W}(o \mid b)$ from unmarked paired counts. For a source-state action law $\widetilde{\mu}_a$, compile

$$\hat{q}_a(o) = \sum_b \widetilde{\mu}_a(b)\,\widehat{W}(o \mid b), \qquad \hat{q}_0(o) = \sum_b \hat{p}(b)\,\widehat{W}(o \mid b), \tag{15}$$

with $\hat{q}_a/\hat{q}_0$ equal to the conditional average of $\widetilde{\mu}_a/\hat{p}$ given the observed state under this fitted categorical model. We obtain $\widetilde{\mu}_a$ by applying Equation (13) to probabilities pooled across contexts. This is an approximation: the order-statistic transform is nonlinear, and coarsening may hide action-dependent differences within a state. Appendix A.4 separates these from transition-estimation error.

Speech: direct regression. Evaluating the action probabilities for each row's source conditional avoids the pooled action-law approximation. The smoothed statewise estimate is

$$\hat{f}_a(o) = \frac{\sum_{j:O_j=o} r_a(X_j) + \alpha}{\#\{j : O_j = o\} + \alpha}, \qquad \alpha > 0, \tag{16}$$

where j indexes paired rows, X_j includes their source context, and O_j is the observed state. Smoothing shrinks sparse-state estimates toward one; without smoothing, the statewise mean minimizes the empirical form of Equation (5). Music uses a fixed 64-state representation; speech selects state count and α using held-out unmarked loss.

4.3 Composite Scoring and Key-Panel Acceptance

The receiver re-encodes the waveform and uses public nonces and session coordinates to associate observed states o_i with sender rows. Each program selects entries from the shared response dictionary to form

$$S(k, m) = \sum_i \log \hat{e}_{\,a_i(k,m)}(o_i), \tag{17}$$

where $\hat{e}$ is $\hat{q}_a$ for compiled emissions or $\hat{f}_a$ for direct responses. The compiled reference term $\sum_i \log \hat{q}_0(o_i)$ is common to all hypotheses on the same audio, so within-audio comparisons use the emissions directly. With a uniform prior over sixteen payloads,

$$A(k) = \log \sum_m \pi(m)\, e^{S(k,m)}, \qquad \hat{m}(k) \in \arg\max_m \big\{ \log \pi(m) + S(k, m) \big\}. \tag{18}$$

Equation (18) implements the owner/decoding split of Section 3.4 with a composite likelihood (Varin et al., 2011): it replaces the observed ratio $\mathbb{E}[\prod_i r_{i,a_i(k,m)} \mid Y]$ of the averaged hypothesis by a product of local responses that pool reference histories and omit across-row dependence. Appendix A.5 analyzes response-estimation error and stability of this ranking rule.

A queried key K_0 is accepted only when $A(K_0)$ strictly exceeds all $n-1$ reference scores; ties reject. Conditional on the audio, public coordinates, and frozen scoring procedure, exchangeability of the queried and reference keys makes the upper rank a valid randomization p-value (Phipson & Smyth, 2010). Unique first place then has probability at most $1/n$, or 10^{-5} for $n = 100{,}000$. The main music experiment instead reports empirical performance on a fixed owner-derived panel (Appendices A.6 and C.4).

Scoring reuses two symbol contributions per row and key across all payloads; Appendix B.6 times 100,000 keys and sixteen payloads.

5 Experiments

We evaluate fixed-sender response learning with transition compilation (music), unmarked receiver selection with direct responses (speech), frozen-dictionary reuse, baselines, and transfer.

Frozen MusicGen-medium (Copet et al., 2023) generates twelve-second sessions of three separate four-second observations carrying the same four-bit payload. Evaluation uses 512 marked sessions

Table 1: Joint owner–payload recovery with the sender fixed, out of 512 twelve-second music sessions. Compositions and NeMo-FSQ are excluded from fitting.

Channel	Identity	Native-only (192 pairs)	Pooled (576 pairs)
Native	512	512	512
DAC	512	512	512
SNAC	475	512	512
DAC → SNAC	385	503	501
SNAC → DAC	379	506	505
NeMo-FSQ	508	512	512

(32 per payload) and 512 unmarked sessions; 256 additional unmarked sessions calibrate presence thresholds. Each source is evaluated through native rendering, DAC (Kumar et al., 2023), SNAC (Siuzdak et al., 2024), both composition orders, and NeMo-FSQ (Langman et al., 2024). These are paired channel views, not independent sessions. Response fitting uses 192 unmarked clips from sixteen disjoint prompts; native, DAC, and SNAC views yield 576 pairs. Both compositions and NeMo-FSQ are excluded from fitting (Appendix C.1).

Joint recovery requires generating-key acceptance and exact payload recovery. *Supplied-key payload recovery* evaluates decoding without owner acceptance. The music panel contains K_0 and 99,999 owner-derived HMAC decoys; results are empirical fixed-panel outcomes. Verification uses the waveform, public nonce, round configuration, and acoustic-frame correspondence. Appendix C details fitting, calibration, and evaluation.

5.1 Does Learning the Response Improve Verification?

Only the transition estimate changes; sender, audio, source action laws, state map, keys, and scoring remain fixed. Identity transport uses $W = I$; native-only fitting uses 192 native pairs; pooled fitting uses all 576 views. Table 1 reports twelve-second recovery; Table 10 adds eight-second results and three count-matched pooled fits.

Pooled responses raise composition recovery from 385 to 501 and from 379 to 505. On DAC-to-SNAC, 117 sessions succeed only with pooled responses and one only with identity (exact McNemar $p = 6 \times 10^{-33}$, Holm-corrected over twelve channel–length tests). Native-only fitting matches or exceeds pooled fitting; count-matched pooled fits recover 500–502 and 504–506 sessions on the two compositions. Transition fitting thus improves recovery without requiring more fitting views or exposure to the evaluated compositions.

Across the four nontrivial channel rows, identity leaves 301 session–channel cases without joint success; 299 nevertheless have the correct supplied-key payload. Pooled responses leave 18 failures, all with correct payloads (Table 11). The gain is primarily in owner acceptance. Decoding compares 16 payloads under the supplied key, whereas acceptance must exceed 99,999 reference scores, so score noise from miscalibrated responses plausibly penalizes attribution far more (Section 3.4).

5.2 Unmarked Receiver Learning and Selection

Speech uses frozen Kyutai TTS (Zeghidour et al., 2025) with Mimi (Défossez et al., 2024) and retains per-row source targets. Twelve candidates vary state count and smoothing. We fit on 72 unmarked observations, select on 48 observations from disjoint source groups, and assess risk on another sixteen held-out groups. Selection chooses 32 states and $\alpha = 1$ before marked evaluation. All candidates predict the same 48 targets with equal weight, covering $J \in \{8, 16\}$ on both active heads; the speech program uses the $J = 8$ subset. Held-out risk decreases from 0.9444 for the neutral response to 0.9137 for the selected receiver (reduction 0.031, 95% interval $[0.026, 0.035]$; Table 15). By Equation (8), the 3.3% relative reduction estimates a lower bound on the weighted fraction of source query χ^2 that survives into the 32-state representation.

Marked evaluation uses 64 sessions of six separate sixteen-second observations (96 seconds total). Recovery is $64/64$ on native rendering, FLAC, and Opus, and $61/64$ on MP3, excluded from fitting and selection. The same-state source-emission control reaches $64/64$ on the first three channels

and $62/64$ on MP3 (Table 23). Held-out risk thus establishes non-neutral prediction, not a recovery advantage over this control. Appendices E.7 and E.2 report budget and sensitivity studies.

A separate neural-codec panel compares channel-specific, pooled, and leave-one-out direct-response fits at matched budgets (Table 2). Unmarked selection chooses 16 states and $\alpha = 1$ for the pooled receiver. Relative to source emission, it adds seven joint successes on SNAC and four on each composition. Held-out SNAC 24 kHz improves from 51 to 57, with seven gains and one regression (Table 14). Every channel has more gains than regressions, but only SNAC reaches uncorrected exact McNemar $p < 0.05$ ($p = 0.016$; 0.125 per composition; 0.070 held out), so these 64-session results indicate direction, not effect size.

Table 2: Speech neural-codec evaluation: joint successes out of 64 sessions, each totaling 96 seconds. Each fit uses 72 fitting and 48 selection observations. The pooled receiver has 16 states; SNAC 24 kHz is excluded from fitting and selection.

Channel	Source emission	Channel-specific	Pooled	Leave-one-out
Native Mimi	64	64	64	64
DAC	63	63	63	63
SNAC	54	60	61	60
DAC → SNAC	53	57	57	55
SNAC → DAC	53	56	57	57
SNAC 24 kHz	51	—	57	—

5.3 Reusing a Frozen Receiver Across Owner Keys

Entries are indexed by action and state, not key, so reuse holds by construction; this experiment checks that results are not specific to K_0's panel. The pooled music dictionary is frozen before evaluating eight new owners, each with sixteen payloads. The 128 marked sessions and 128 controls hold the 64-prompt set fixed across owner keys. Joint recovery is $128/128$ on four channels and $127/128$ on both compositions; every supplied-key payload is recovered. No control is accepted under either the single-key or any-of-eight test (Table 13).

Separate null checks yield two unique-first events in 393,216 dependent wrong-key queries, one false attribution among 640 source-disjoint controls, and none among 4,096 FMA-small recordings (Defferrard et al., 2017). Source-disjoint marked recovery is $639/640$ (Tables 21 and 22). These finite empirical checks do not establish a 10^{-5} false-acceptance rate.

5.4 Comparison with Existing Audio Watermarks

Aligned-IS and HPT are keyed generation-time baselines; AudioSeal is a post-hoc method with message-based attribution (Wu et al., 2025b; Milis et al., 2026; San Roman et al., 2024). We use their official implementations on common prompts, channel paths, and twelve-second budgets. The evaluated AudioSeal configuration participates in presence detection, not CHORD's secret-key panel comparison. HPT uses $h = 0$, $\gamma = 0.25$, and $\delta = 1$ on heads 0–3, frozen Leiden maps, and no codec fine-tuning (Appendix C.6).

Key identification. Under the common unique-first rule, CHORD recovers owner and payload in 501–512 sessions per channel (Table 3). Aligned-IS identifies the key in 510 native and 498 DAC sessions, but 7 and 4 under the compositions. HPT never places the generating key uniquely first. On native audio, its median conservative rank is 5,571, with no generating-key tie for first (Appendix D.2). This failure concerns identification among competing keys, not the absence of a presence signal.

Calibrated presence. On SNAC and the compositions, HPT detects 61–123 marked sessions versus CHORD's 13–33; CHORD is not uniformly better on this endpoint; its raw-score AUC is 0.949–0.975. Table 3's unique-first rule is itself a presence decision, with nominal level $1/n$ under exchangeable keys. Table 4 instead thresholds the raw payload log-mean-exp $A(k)$ without reference subtraction; the reference term omitted in Section 4.3 cancels between keys on the same audio but not across recordings, which lowers power on the harder channels.

Table 3: Unique-first identification over 100,000 keys on the fixed key panel (Appendix C.4). Successes out of 512 twelve-second marked sessions; false acceptances out of 512 controls in parentheses. CHORD entries also require exact four-bit payload recovery.

Channel	Aligned-IS	HPT	CHORD (identity)	CHORD (learned)
Native	510 (0)	0 (0)	512 (0)	**512** (0)
DAC	498 (0)	0 (0)	512 (0)	**512** (0)
SNAC	12 (0)	0 (0)	475 (0)	**512** (0)
DAC → SNAC	7 (0)	0 (0)	385 (0)	**501** (0)
SNAC → DAC	4 (0)	0 (0)	379 (0)	**505** (0)
NeMo-FSQ	294 (0)	0 (0)	508 (0)	**512** (0)

Table 4: Calibrated presence: true positives out of 512 marked sessions, with test-set false positives out of 512 controls in parentheses. Thresholds use 256 independent controls at target FPR 10^{-2}.

Channel	Aligned-IS	HPT	AudioSeal	CHORD (score)
Native	503 (1)	464 (3)	512 (7)	511 (1)
DAC	281 (1)	384 (2)	512 (2)	474 (2)
SNAC	0 (1)	123 (5)	13 (3)	33 (1)
DAC → SNAC	0 (1)	99 (3)	17 (2)	13 (1)
SNAC → DAC	0 (2)	61 (1)	25 (5)	25 (2)
NeMo-FSQ	2 (2)	116 (1)	62 (0)	116 (1)

5.5 Quality and Limits of Transfer

Paired CLAP evaluation (Wu et al., 2023) gives a CHORD–unmarked difference of -0.0013 (95% interval $[-0.0056, 0.0029]$; 512 sessions). Speech word- and character-error differences also have intervals containing zero (Table 17), indicating no detectable difference in these metrics rather than perceptual equivalence.

On music, SpeechTokenizer and Mimi yield $9/16$ joint successes despite $16/16$ supplied-key payload recovery (Table 28). In the 64-session removal panel, polarity inversion, 10 ms delay, and $1.1\times$ speed give zero fixed-rule successes but 64 after search. Time stretching and pitch shifting yield $1/64$; both DCCRN conditions yield zero (Hu et al., 2020); Table 16 gives search budgets. These are failures of the evaluated receiver, not evidence-erasure tests. The selected speech receiver recovers only $4/64$ native sessions from a twelve-second prefix, versus $64/64$ over six observations (96 s; Table 24).

Limitations. Verification assumes a known public nonce, round configuration, and acoustic-frame correspondence; locating them for audio of unknown origin adds looks whose cost enters the LR/n bound (Appendix A.6). Synchronization edits and enhancement defeat the fixed-rule receiver, and the payload has four bits. The 10^{-5} guarantee assumes exchangeable randomized keys, whereas our panels are fixed and owner-derived. Each modality uses one frozen generator, and music compilation relies on the pooled action-law approximation of Section 4.2.

6 Conclusion

We introduced CHORD, a CHannel-Observed Response Dictionary for verifying generation-time audio watermarks after re-encoding. Rather than making symbols invariant to retokenization, CHORD learns how the channel transforms a known sender's evidence: observed ratios are conditional expectations of source ratios, so unmarked pairs suffice to fit and select ordered-candidate rank-action responses shared by all keys and payloads. With the sender fixed, these responses raise music recovery through codec compositions from about 75% to 98%, mainly by restoring owner attribution, and the frozen dictionary transfers to new owners. Future work targets synchronization, larger payloads, and other distribution-preserving senders.

AI USE STATEMENT

We used ChatGPT, Codex, and Claude solely to assist with manuscript drafting, editing, and internal review. The authors reviewed and revised all AI-assisted text and take responsibility for the final content. The experimental use of MusicGen and Kyutai TTS is documented in Section 5 and Appendix B.

ETHICS STATEMENT

CHORD aims to support provenance verification for generated audio. Watermark decisions can produce false attributions or missed detections and should be interpreted as statistical evidence rather than definitive proof of ownership or authenticity. We report false-attribution controls and robustness limitations, and recommend considering watermark evidence alongside other provenance information.

REPRODUCIBILITY STATEMENT

Supporting derivations are provided in Appendix A, and most implementation details are documented in Appendix B. Appendix C describes the data splits, receiver selection, and evaluation protocols, while Appendices D and E provide detailed results and additional experiments. We plan to release the code publicly after the review process.

REFERENCES

Scott Aaronson. My AI safety lecture for UT effective altruism. Shtetl-Optimized blog, November 2022. URL https://scottaaronson.blog/?p=6823. November 28, 2022.

Johann Brehmer, Gilles Louppe, Juan Pavez, and Kyle Cranmer. Mining gold from implicit models to improve likelihood-free inference. *Proceedings of the National Academy of Sciences*, 117(10):5242–5249, 2020. doi: 10.1073/pnas.1915980117.

Guangyu Chen, Yu Wu, Shujie Liu, Tao Liu, Xiaoyong Du, and Furu Wei. WavMark: Watermarking for audio generation. *arXiv preprint arXiv:2308.12770*, 2023. doi: 10.48550/arXiv.2308.12770. URL https://arxiv.org/abs/2308.12770.

Yen-Shan Chen, Shih-Yu Lai, Ying-Jung Tsou, Yi-Cheng Lin, Bing-Yu Chen, Yun-Nung Chen, Hung-yi Lee, and Shang-Tse Chen. Latent-Mark: An audio watermark robust to neural codec compression. *arXiv preprint arXiv:2603.05310*, 2026. doi: 10.48550/arXiv.2603.05310. URL https://arxiv.org/abs/2603.05310v3. Version 3; Interspeech 2026, long paper track.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 47704–47720, 2023. doi: 10.52202/075280-2066. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/94b472a1842cd7c56dcb125fb2765fbd-Abstract-Conference.html.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A dataset for music analysis. In *Proceedings of the 18th International Society for Music Information Retrieval Conference*, pp. 316–323, 2017. URL https://archives.ismir.net/ismir2017/paper/000075.pdf.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: A speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024. doi: 10.48550/arXiv.2410.00037. URL https://arxiv.org/abs/2410.00037v2.

Steffen Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimiliano Pontil. Conditional mean embeddings as regressors. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 1803–1810. Omnipress, 2012. URL https://icml.cc/2012/papers/898.pdf.

Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. In *Interspeech 2020*, pp. 2472–2476, 2020. doi: 10.21437/Interspeech.2020-2537. URL https://www.isca-archive.org/interspeech_2020/hu20g_interspeech.html.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. Unbiased watermark for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=uWVC5FVidc.

Nikola Jovanović, Ismail Labiad, Tomáš Souček, Martin Vechev, and Pierre Fernandez. Watermarking autoregressive image generation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2506.16349.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17061–17084. PMLR, 2023. URL https://proceedings.mlr.press/v202/kirchenbauer23a.html.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=FpaCL1MO2C.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved RVQGAN. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27980–27993, 2023. doi: 10.52202/075280-1214. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/58d0e78cf042af5876e12661087bea12-Abstract-Conference.html.

Ryan Langman, Ante Jukić, Kunal Dhawan, Nithin Rao Koluguri, and Boris Ginsburg. Spectral codecs: Spectrogram-based audio codecs for high quality speech synthesis. *arXiv preprint arXiv:2406.05298*, 2024. doi: 10.48550/arXiv.2406.05298. URL https://arxiv.org/abs/2406.05298v1. Version 1.

Yixin Liu, Lie Lu, Jihui Jin, Lichao Sun, and Andrea Fanelli. XAttnMark: Learning robust audio watermarking with cross-attention. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 38987–39015. PMLR, 2025. URL https://proceedings.mlr.press/v267/liu25ap.html.

Georgios Milis, Yubin Qin, Yihan Wu, and Heng Huang. Hidden in plain tokens: Simply robust, gradient-free watermark for synthetic audio. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2605.25967.

Patrick O'Reilly, Zeyu Jin, Jiaqi Su, and Bryan Pardo. Deep audio watermarks are shallow: Limitations of post-hoc watermarking techniques for speech. In *The 1st Workshop on GenAI Watermarking*, 2025. URL https://arxiv.org/abs/2504.10782. Collocated with ICLR 2025.

Yigitcan Özer, Woosung Choi, Joan Serrà, Mayank Kumar Singh, Wei-Hsiang Liao, and Yuki Mitsufuji. A comprehensive real-world assessment of audio watermarking algorithms: Will they survive neural codecs? In *Interspeech 2025*, pp. 5113–5117, 2025. doi: 10.21437/Interspeech.2025-1530. URL https://www.isca-archive.org/interspeech_2025/ozer25_interspeech.html.

Belinda Phipson and Gordon K. Smyth. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1): Article 39, 2010. doi: 10.2202/1544-6115.1585.

Robin San Roman, Pierre Fernandez, Hady Elsahar, Alexandre Défossez, Teddy Furon, and Tuan Tran. Proactive detection of voice cloning with localized watermarking. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43180–43196. PMLR, 2024. URL https://proceedings.mlr.press/v235/san-roman24a.html.

Hubert Siuzdak, Florian Grötschla, and Luca A. Lanzendörfer. SNAC: Multi-scale neural audio codec. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024. URL https://openreview.net/forum?id=PFBF5ctj4X.

Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42, 2011. URL https://www3.stat.sinica.edu.tw/statistica/oldpdf/A21n11.pdf.

Yihan Wu, Zhengmian Hu, Junfeng Guo, Hongyang Zhang, and Heng Huang. A resilient and accessible distribution-preserving watermark for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53443–53470. PMLR, 2024. URL https://proceedings.mlr.press/v235/wu24h.html.

Yihan Wu, Ruibo Chen, Georgios Milis, Junfeng Guo, and Heng Huang. A watermark for auto-regressive speech generation models. In *Interspeech 2025*, pp. 3474–3478, 2025a. doi: 10.21437/Interspeech.2025-2328. URL https://www.isca-archive.org/interspeech_2025/wu25k_interspeech.html.

Yihan Wu, Georgios Milis, Ruibo Chen, and Heng Huang. Robust distortion-free watermark for autoregressive audio generation models. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://arxiv.org/abs/2510.21115.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. URL https://arxiv.org/abs/2211.06687v3.

Neil Zeghidour, Eugene Kharitonov, Manu Orsini, Václav Volhejn, Gabriel de Marmiesse, Edouard Grave, Patrick Pérez, Laurent Mazaré, and Alexandre Défossez. Streaming sequence-to-sequence learning with delayed streams modeling. *arXiv preprint arXiv:2509.08753*, 2025. doi: 10.48550/arXiv.2509.08753. URL https://arxiv.org/abs/2509.08753v2.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. SpeechTokenizer: Unified speech tokenizer for speech language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AF9Q8Vip84.

A SUPPORTING DERIVATIONS

We use the notation of Sections 3–4: P is the unmarked source law, $\nu(dx, dy) = P(dx)T(dy \mid x)$ is the unmarked joint law, $Q = T_{\#}P$ is its observation marginal, and $\mathcal{R}h = \mathbb{E}_\nu[h(X) \mid Y]$. All expectations below are under this joint law unless another measure is specified. The response identities concern the chosen source queries; the score analysis in Appendix A.5 concerns the local composite rule in Equation (17).

A.1 THE FAMILY-LEVEL REDUCTION

Assume standard Borel source and observation spaces and fix the public conditioning information. Let $P_a \ll P$ be any source alternative with ratio $r_a = dP_a/dP$ and observation law $Q_a = T_{\#}P_a$.

Proof of Equation (3) *and Proposition 1.* For every bounded measurable φ,

$$\begin{aligned}\int \varphi(y)Q_a(dy) &= \int\int \varphi(y)T(dy \mid x)r_a(x)P(dx)\\ &= \mathbb{E}_\nu[r_a(X)\varphi(Y)]\\ &= \int \varphi(y)\mathbb{E}_\nu[r_a(X) \mid Y = y]Q(dy). \end{aligned} \tag{19}$$

Thus $Q_a \ll Q$ and $dQ_a/dQ = \mathbb{E}_\nu[r_a(X) \mid Y]$, Q-almost surely. Taking $r_a = w_{m,k}$ proves Equation (3); subtracting one gives $Q_{m,k}/Q = 1 + \mathcal{R}h_{m,k}$.

For $h \in L^2(P)$, conditional Jensen's inequality gives

$$\|\mathcal{R}h\|^2_{L^2(Q)} \le \|h\|^2_{L^2(P)}. \tag{20}$$

The operator is linear and continuous. Consequently, knowing its action on every $h_{m,k}$ determines its action on their span $\mathcal{H}$ and on the closure of that span. Two channels with the same P and Q have the same family of observation ratios if and only if their responses agree on each $h_{m,k}$, equivalently on $\mathcal{H}$. Equality is understood up to Q-null sets. □

The pairing example in Section 3.2 is an identification argument. Independent samples from the two marginals have identical laws under $Y = X$ and $Y = -X$, while the response to $h(x) = \rho x$ has opposite sign. Thus unpaired marginal samples do not identify the response over a class containing both channels.

A.2 PREDICTION RISK, REPRESENTATION, AND EVIDENCE

Write $f^\star_a(Y) = \mathbb{E}_\nu[r_a(X) \mid Y]$ and use $L^2(Q)$ norms for functions of Y.

Prediction and representation identities. For one query, decompose

$$r_a(X) - f_a(Y) = \big(r_a(X) - f^\star_a(Y)\big) + \big(f^\star_a(Y) - f_a(Y)\big). \tag{21}$$

The first term has conditional mean zero given Y and is orthogonal to the second. Therefore

$$\mathbb{E}_\nu(r_a - f_a)^2 = \mathbb{E}_\nu(r_a - f^\star_a)^2 + \|f_a - f^\star_a\|^2_2. \tag{22}$$

Summing with the fixed weights λ_a proves Equation (6). For $Z = g(Y)$, the tower property gives $f^\star_{a,Z} = \mathbb{E}_\nu[f^\star_a \mid Z]$. Orthogonality to functions of Z then yields

$$\|f^\star_a - f_{Z,a}\|^2_2 = \|f^\star_a - f^\star_{a,Z}\|^2_2 + \|f^\star_{a,Z} - f_{Z,a}\|^2_2, \tag{23}$$

which proves Equation (7). □

Since $f^\star_a = dQ_a/dQ$, we have $\|f^\star_a - 1\|^2_2 = \chi^2(Q_a\|Q)$. Subtracting the excess-risk identities for f and the neutral predictor gives Equation (8).

Surviving-evidence lower bound. Because $\mathbb{E}_P r_a = 1$, the neutral risk is $L(\mathbf{1}) = \sum_a \lambda_a \mathbb{E}_P(r_a - 1)^2 = \sum_a \lambda_a \chi^2(r_a P \,\|\, P)$. For a representation $Z = g(Y)$, Equations (7) and (8) give $L(\mathbf{1}) - L(f_Z) = \sum_a \lambda_a [\chi^2(Q_a^Z \| Q^Z) - \|f^\star_{a,Z} - f_{Z,a}\|_2^2]$, where Q_a^Z and Q^Z are the laws of Z. The data-processing inequality for f-divergences then yields

$$\frac{L(\mathbf{1}) - L(f_Z)}{L(\mathbf{1})} \le \frac{\sum_a \lambda_a \chi^2(Q_a^Z \| Q^Z)}{\sum_a \lambda_a \chi^2(r_a P \| P)} \le \frac{\sum_a \lambda_a \chi^2(Q_a \| Q)}{\sum_a \lambda_a \chi^2(r_a P \| P)} \le 1. \tag{24}$$

With f_Z fitted on separate data, held-out losses estimate the left side without bias in each risk, so the relative risk reduction estimates a lower bound on the weighted fraction of query distinguishability that survives the channel and the representation. It is not an estimate of that fraction itself, because estimation error and discarded information both reduce it.

In particular,

$$\|f_a - f_a^\star\|_2^2 \le \frac{L(f) - L(f^\star)}{\lambda_a}, \qquad |\mathbb{E}_Q f_a - 1|^2 \le \|f_a - f_a^\star\|_2^2. \tag{25}$$

The second inequality uses $\mathbb{E}_Q f_a^\star = 1$. Fixed query weights therefore determine how aggregate prediction error controls individual responses. Comparing representation candidates requires the same source queries and weights; changing the receiver state map does not change these targets.

Owner and payload components. Equation (9) gives $h_{m,k} = u_k + v_{m,k}$ with $\sum_m \pi(m) v_{m,k} = 0$. Linearity of $\mathcal{R}$ proves Equation (10). Expanding the squared norm of $\mathcal{R}u_k + \mathcal{R}v_{m,k}$ and averaging over m cancels the cross term, proving Equation (11). The empirical outcome decomposition in Table 11 distinguishes acceptance from supplied-key decoding; it does not estimate these squared norms.

A.3 Ordered-candidate probabilities

Proof of Proposition 2. Fix a history and candidate occurrences $(Z_0, \dots, Z_{J-1})$. Sorting permutes the occurrences, including duplicates. Under the independent uniform phase in the proposition, the payload shift preserves a uniform rank U, so for any token set D,

$$\mathbb{P}(Z_{(U)} \in D \mid Z_0, \dots, Z_{J-1}) = \frac{1}{J} \sum_{j=0}^{J-1} \mathbf{1}\{Z_j \in D\}. \tag{26}$$

Averaging candidates drawn from p_c returns $p_c(D)$. A fresh phase independent of the history gives this equality at each autoregressive step, establishing the reference law for every fixed payload under the proposition's randomization. □

For a fixed zero-indexed rank u, the event $Z_{(u)} \le x$ occurs when at least $u + 1$ of the J independent draws are at most x. Its probability is $H_{J,u}(F_c(x))$. Taking the difference across the atom at x gives Equation (13). Moreover,

$$0 \le p_c^{J,u}(x) \le 1 - (1 - p_c(x))^J \le J p_c(x), \tag{27}$$

proving the bounded local-ratio property on the support of p_c.

Averaging over ranks returns the reference law. In the fitted categorical model and for the direct targets, respectively,

$$\frac{1}{J}\sum_u \widetilde{\mu}_{J,u,h} = \widehat{p}_h, \qquad \frac{1}{J}\sum_u \widehat{q}_{J,u,h} = \widehat{p}_h \widehat{W}_h, \qquad \frac{1}{J}\sum_u r_{J,u}(X) = 1. \tag{28}$$

Equation (16) preserves the last identity at each observed state because all ranks use the same observations and smoothing mass. If $\mathbb{E}_K P_{m,K} = P$, applying a key-independent channel also gives $\mathbb{E}_K Q_{m,K} = Q$. These are properties of the stated randomized model; the candidate and phase execution used in each system is specified in Section 4.1 and Appendix B.1.

A.4 FINITE-STATE RESPONSE ESTIMATION

For a chosen query r_a, let $P_a = r_a P$, $B = b_X(X)$ be the source state, and $O = b_Y(Y)$ the observed state. Write $\mu_a(b) = P_a(B = b)$, $W_0(o \mid b) = \mathbb{P}_\nu(O = o \mid B = b)$, and $W_a(o \mid b) = \mathbb{P}_{r_a\nu}(O = o \mid B = b)$. Then the query's observed-state marginal is $q_a^{\mathrm{loc}} = \mu_a W_a$.

Compilation. The pooled action law $\widetilde{\mu}_a$ of Equation (15) differs from the actual query marginal μ_a here. The music construction is $\widehat{q}_a = \widetilde{\mu}_a \widehat{W}_0$. For $\widehat{q}_0(o) > 0$, its fitted-model ratio is

$$\frac{\widehat{q}_a(o)}{\widehat{q}_0(o)} = \sum_{b:\widehat{p}(b)>0} \frac{\widetilde{\mu}_a(b)}{\widehat{p}(b)} \frac{\widehat{p}(b)\widehat{W}_0(o \mid b)}{\widehat{q}_0(o)}. \tag{29}$$

The second factor is the fitted source-state posterior. Thus compilation is exactly a conditional average for that categorical model. The approximation to the original query can be separated as

$$\begin{aligned} \|\widehat{q}_a - q_a^{\mathrm{loc}}\|_1 \le \|\widetilde{\mu}_a - \mu_a\|_1 &+ \sum_b \mu_a(b) \|\widehat{W}_0(\cdot \mid b) - W_0(\cdot \mid b)\|_1 \\ &+ \sum_b \mu_a(b) \|W_0(\cdot \mid b) - W_a(\cdot \mid b)\|_1. \end{aligned} \tag{30}$$

Insert $\mu_a \widehat{W}_0$ and $\mu_a W_0$ and use the triangle inequality and contraction by a stochastic matrix. The terms are pooled action-law error, transition estimation error, and within-state query variation. For the last term, let $\bar{r}_a(B) = \mathbb{E}_\nu[r_a(X) \mid B]$. Since

$$\begin{aligned} \mu_a(b) W_a(o \mid b) &= \mathbb{E}_\nu[r_a(X) \mathbf{1}\{B = b, O = o\}], \\ \mu_a(b) W_0(o \mid b) &= \mathbb{E}_\nu[\bar{r}_a(B) \mathbf{1}\{B = b, O = o\}], \end{aligned}$$

summing absolute differences bounds it by $\mathbb{E}_\nu |r_a(X) - \bar{r}_a(B)|$. Zero-mass conditional states have zero weight. In music, pooling before the nonlinear order-statistic transform affects the first term; learning $\widehat{W}_0$ addresses the second.

Direct regression. Speech instead evaluates the per-row query before grouping examples by observed state. Without smoothing, the statewise mean minimizes empirical squared loss for those fixed targets. Equation (16) adds mass α at the neutral value one, shrinking sparse-state estimates without changing the target dictionary. Representation and smoothing candidates can therefore be compared on the common loss in Equation (5).

A.5 LOCAL SCORE COMPOSITION AND STABILITY

The receiver in Section 4.3 evaluates the local score in Equation (17), followed by payload marginalization in Equation (18). Three issues are distinct. *Local response error* concerns estimated action responses, as in Appendix A.4. *Context pooling* replaces history-specific responses by a shared dictionary. *Across-row dependence* is not modeled by adding local log responses. The following statement controls perturbations of this specified composite score; it does not identify the score with a full received-sequence likelihood.

Ranking stability. Let $e^\star_{i,a}$ and $\widehat{e}_{i,a}$ be positive target and estimated entries on the scored support. Define $S^\star$ and S by the same action programs and Equation (17). If

$$\sup_{a,o} |\log \widehat{e}_{i,a}(o) - \log e^\star_{i,a}(o)| \le \varepsilon_i, \tag{31}$$

then the triangle inequality gives

$$\sup_{k,m} |S(k,m) - S^\star(k,m)| \le \sum_i \varepsilon_i = \delta. \tag{32}$$

Writing $A^\star$ for the corresponding payload-aggregated score, exponentiation and summation imply

$$e^{-\delta} \sum_m \pi(m) e^{S^\star(k,m)} \le \sum_m \pi(m) e^{S(k,m)} \le e^{\delta} \sum_m \pi(m) e^{S^\star(k,m)},$$

so $|A(k) - A^\star(k)| \le \delta$. A target owner gap, or a target within-owner payload gap, greater than 2δ therefore preserves its strict ordering. If both entries are at least $b > 0$ and their uniform difference is at most η_i, the mean-value theorem gives $\varepsilon_i \le \eta_i / b$.

Equation (5) measures average squared prediction error under the unmarked law, whereas the ranking-stability result uses pointwise log-error bounds and a score margin. The experiments therefore evaluate prediction risk and composite decisions separately.

A.6 Reference-key ranks and repeated queries

Condition on the observation, public coordinates, and frozen common score function, collected in C. If $(K_0, \dots, K_{n-1})$ are exchangeable conditional on C, the conservative normalized rank

$$p_{\text{key}} = \frac{1 + \sum_{j=1}^{n-1} \mathbf{1}\{A(K_j) \ge A(K_0)\}}{n} \tag{33}$$

satisfies $\mathbb{P}(p_{\text{key}} \le \alpha \mid C) \le \alpha$ for $0 \le \alpha \le 1$.

Condition further on the multiset of scores. Exchangeability assigns the queried label uniformly among its positions. At most $\lfloor n\alpha \rfloor$ positions have conservative rank at most $n\alpha$; assigning each tie its largest rank cannot increase this count. Averaging proves the claim (Phipson & Smyth, 2010). Unique first place consequently has probability at most $1/n$. The probability includes key randomization, and does not follow merely from a key's failure to generate the audio.

For L queried keys and R fixed looks, a union bound gives LR/n when every query–look pair satisfies the same conditional exchangeability assumptions. No independence between looks is required. If s labels in a common panel are queried on each of M inputs, at most one label can be uniquely first on each input, and

$$\mathbb{E}[\text{total unique-top events}] \le \frac{Ms}{n}. \tag{34}$$

Equality holds under exchangeability and an almost surely unique maximum. Linearity of expectation does not require independent inputs or channel views. Appendix C.4 distinguishes this randomization argument from the empirical fixed-key panels.

B Implementation

This appendix specifies the two implementations in Section 4. Music compiles a fixed-state transition table, whereas speech regresses directly on row-level action targets. In both cases the learned response dictionary is indexed by action and observed state, not by owner or payload. Generation traces are needed for fitting; verification uses the received waveform and public session metadata.

B.1 Generators and source sampling

Table 5: **Audio configurations.** Heads are indexed from zero. All model weights remain fixed.

Setting	Music	Speech
Generator	MusicGen-medium	Kyutai TTS 0.75B
Native audio rate	32 kHz	24 kHz
Native tokenizer	MusicGen codec	Mimi
Codebooks / vocabulary	4 / 2,048	16 / 2,048
Acoustic frame rate	50 Hz	12.5 Hz
Active heads	0, 1	0, 1
Sampling temperature	1.0	0.6
Top-k	250	250
Frames per observation	200	200
Observations per session	3	6
Candidate counts	8, 8, 16	8 in each observation
Payloads	16	16

Music uses `facebook/musicgen-medium`, with classifier-free guidance 3.0. Speech uses `kyutai/tts-0.75b-en-public`. Source probabilities are truncated, renormalized, and converted by largest-remainder allocation to integer counts totaling 2^{52}. Inverse-CDF sampling on this grid defines the implemented reference policy. Music model evaluation uses FP32; speech uses BF16 weights and FP32 softmax. Probability processing and receiver scores use FP64.

Candidate randomness and keyed execution. The music execution follows Section 4.1: candidate draws and rank phases are obtained from the same owner key and nonce through domain-separated HMAC-SHA256 inputs containing the observation round, acoustic frame, and head. The receiver constructs its action dictionary using the independent-draw probabilities in Equation (13). These probabilities define the receiver's action queries; fixing a music key, coordinates, and history fixes the executed candidate draws.

Speech uses a seeded candidate stream separate from the owner key and evaluates the action targets from each row's source conditional. Its HMAC phase domains also include the model identifier. Proposition 2 states the ideal independent-phase result; execution uses pseudorandom phases. Each received observation is supplied with its nonce and round configuration, but not its generation prompt.

B.2 Ordering, states, and payload schedule

For music, an independent permutation of 2,048 equally spaced public scalar values defines each head's token ordering. The evaluated ordering uses NumPy `default_rng(659009)`. Candidate occurrences are sorted by scalar, with draw index breaking ties. Writing $\rho_h(x) \in \{0, \dots, 2047\}$ for the token's position in this order, the music state map is

$$b_h(x) = \left\lfloor \frac{64\rho_h(x)}{2048} \right\rfloor. \tag{35}$$

Each of the 64 bins contains 32 tokens.

For speech, tokens are ordered by SHA256 digests of the model identifier, head, and token identifier, with token index breaking ties. A candidate representation with B states uses $b_{h,B}(x) = \lfloor B\rho_h(x)/2048 \rfloor$. Both sides of a source-emission control use the same map.

The four-bit payload uses the fifteen nonzero binary characters. For character j and message m,

$$z_j(m) = (-1)^{\mathrm{popcount}(m \& j)}. \tag{36}$$

In music, characters 1 through 15 occupy consecutive frame blocks. Characters $1, 2, 4, 8, 15$ receive fourteen frames and the others thirteen, totaling 200. In speech the schedule is cyclic, $j_t = 1 + (t \bmod 15)$. A positive symbol adds $J/2$ to the keyed rank; a negative symbol adds zero. The active heads share the payload symbols and use different phases.

B.3 Response estimation

Music fitting uses twelve unmarked clips for each of sixteen prompts. Native, DAC, and SNAC views give 576 pairs from 192 source clips. Counts pool corresponding acoustic frames within each active head: $C_h(b, o)$ counts a source-state/observed-state pair, $n_h(b) = \sum_o C_h(b, o)$, and $N_h = \sum_b n_h(b)$. The marginals are

$$\widehat{p}_h(b) = \frac{n_h(b)}{N_h}, \qquad \widehat{r}_h(o) = \frac{\sum_b C_h(b, o)}{N_h}. \tag{37}$$

The transition estimate uses the observation marginal as a prior of total pseudocount mass one:

$$\widehat{W}_h(o \mid b) = \frac{C_h(b, o) + \widehat{r}_h(o)}{n_h(b) + 1}. \tag{38}$$

For $\widehat{r}_h(o) > 0$, the transition and each normalized compiled action emission are positive. The compiled receiver raises an error for empty states and applies no additional ϵ or floor during scoring. The music receiver uses 64 states; its numerical response estimates use only the unmarked fitting pairs.

The pooled order-statistic law is evaluated using a nonnegative multinomial expression. For bin b, write $l = \widehat{F}_h(b-1)$, $e = \widehat{p}_h(b)$, and $v = 1 - \widehat{F}_h(b)$. Then

$$\widetilde{\mu}_{J,u,h}(b) = \sum_{d=0}^{u} \sum_{a=u+1-d}^{J-d} \binom{J}{d} \binom{J-d}{a} l^d e^a v^{J-d-a}. \tag{39}$$

Here d candidates are below the bin and a are in it. The summation places rank u inside the bin. Laws are row-normalized before multiplication by $\widehat{W}_h$.

The speech trace retains top-250 token identifiers, reference-grid probabilities, sampled token, head, and frame. The oracle computes the selected token's rank-action probability and divides by its reference probability. The target dictionary contains $J = 8$ and $J = 16$ on both active heads. The speech program uses the $J = 8$ responses. Selection assigns weight $1/48$ to each action in the full dictionary: sixteen targets for $J = 8$ and thirty-two for $J = 16$. The selection objective thus covers the full reusable dictionary, including responses beyond the deployed speech program.

B.4 Frame correspondence and score computation

Music's delayed generation index is converted to acoustic frame by $t = L - 1 - h$, where L is the current prefix length and h the head. Re-tokenization uses the native encoder and low-level quantizer at parameter 6.0, yielding four streams of 200 frames. For speech, the native hook corrects the head delays, the TTS delay prefix is removed, and the first 200 re-encoded Mimi frames are used. Both receivers use fixed acoustic-frame correspondence.

At each row and key, the implementation computes the two possible symbol contributions,

$$\ell_i^- = \log \widehat{e}_{J,U_i,h}(o_i), \qquad \ell_i^+ = \log \widehat{e}_{J,(U_i+J/2) \bmod J,h}(o_i). \tag{40}$$

All payload scores reuse them:

$$S(k,m) = \sum_i \ell_i^- + \sum_i \mathbf{1}\{z_i(m) = +1\}(\ell_i^+ - \ell_i^-). \tag{41}$$

This equals eq. (17). Stable log-sum-exp computes $A(k)$, including the uniform-payload offset $-\log 16$. The music presence comparison thresholds this same log-mean-exp of accumulated log emissions without reference subtraction. Music treats reference-score equality conservatively and resolves payload ties by a fixed first maximizer. Speech treats differences within 10^{-10} as ties and requires a unique payload within the same tolerance.

B.5 Source-emission and projected speech controls

Let $R_a(b)$ be the regularized mean of raw-token target $r_a(X)$ within source state b, using the same smoothing formula as Equation (16). The source-emission control evaluates $R_a(o)$ at the observed state. The hard and soft posterior controls also estimate paired counts $C(b,o)$ and

$$\widehat{\Pi}(b \mid o) = \frac{C(b,o)}{\sum_v C(v,o)}. \tag{42}$$

The hard response is $R_a(\arg\max_b C(b,o))$, choosing the lowest state on a tie. The soft response is $\sum_b R_a(b)\widehat{\Pi}(b \mid o)$. Empty observed states receive the neutral response one. A raw-token control uses 2,048 source states. These are receiver substitutions on the same sender, observed audio, and hypothesis program. In particular, source emission evaluates a source-state response at the received state without learning the observed-state response directly.

B.6 Verification-kernel runtime

Table 6 reports a synthetic CPU benchmark with 100,000 keys and sixteen payloads. The scoring kernel takes 7.182 seconds; including phase generation gives 17.058 seconds. Both timings exclude waveform encoding and initialization. They describe the measured scoring components, not end-to-end verification latency or a hardware-independent speed comparison.

Table 6: Synthetic CPU verification benchmark for 100,000 keys and sixteen payloads. Waveform encoding and initialization are excluded.

Measured computation	Time (s)
Scoring kernel	7.182
Phase generation and scoring	17.058

C Evaluation Protocols

A *session* aggregates separately generated observations carrying one payload; a transformed channel view is not an additional independent source session. Fitting counts refer to source–observation pairs, so several views may share one source clip. Session counts and receiver configurations are reported separately for the main music benchmark, supplementary music panels, and speech studies.

C.1 Main music benchmark

The main comparison uses 512 marked sessions, 512 unmarked controls, and an independent 256-session unmarked calibration set. Each marked session comprises three four-second observations carrying a common four-bit payload; controls use the same observation budget. The sixteen payloads have 32 marked sessions each. Each source is evaluated under native rendering, DAC, SNAC, both composition orders, and NeMo-FSQ. The music benchmark uses a fixed 64-state representation, with response values fitted from unmarked pairs. Table 7 consolidates these roles.

Table 7: Main music benchmark. Fitting, calibration, and evaluation have distinct roles; the 576 fitting views come from 192 unmarked clips.

Component	Configuration
Evaluation	512 marked sessions and 512 unmarked controls
Observation budget	Three four-second observations per session
Payload allocation	Sixteen payloads, 32 marked sessions each
Calibration	256 independent unmarked sessions; target FPR 10^{-2}
Response fitting	192 unmarked clips, native/DAC/SNAC views: 576 pairs
Music receiver	64 states, marginal prior mass one; fixed after fitting
Fitting exclusions	Both codec compositions and NeMo-FSQ
Main key panel	Owner key plus 99,999 owner-derived decoys
Verification inputs	Waveform, public nonce, round configuration, frame correspondence

Thresholds in Table 4 are fixed from the independent calibration set, at a common target FPR of 10^{-2}. CHORD uses the raw statistic $A(k)$, the log-mean-exp of the sixteen accumulated payload log-emission scores, without reference subtraction. Key-panel decisions instead compare scores within each audio input and reject ties. The target FPR is a calibration objective; Table 4 reports the realized false positives on separate controls. The supplementary nominal decision rules below do not define the main calibrated thresholds. The key construction is specified in Appendix C.4.

C.2 Supplementary music sources and processing

The supplementary music benchmark uses 64 prompt-indexed sessions. Each session contains three separately generated observations carrying one payload; the session index modulo sixteen assigns balanced payloads. Marked and unmarked generation use different seeds. Each generated source is processed through all six channel paths, so channel results are paired measurements of the same source sessions.

The 44.1-kHz paths resample from and back to the host rate. Compositions retain a 32-kHz PCM16 waveform between codecs. CHORD's common-benchmark audio is generated in CPU FP32; Aligned-IS uses CUDA FP32 with TF32 disabled. AudioSeal marks the unmarked host audio. All methods share prompts and processing paths.

Table 8: **Music processing settings.** All receiver inputs return to the 32-kHz host representation.

Component	Setting
Host waveform	Four seconds, 32 kHz, PCM16
DAC	44.1-kHz model, pretrained 0.0.1 weights, 8 kbit/s
SNAC	`hubertsiuzdak/snac_32khz`
NeMo-FSQ	`nvidia/audio-codec-44khz`, approximately 6.9 kbit/s
Compositions	DAC then SNAC; SNAC then DAC
Received length	128,000 samples, cropped or padded after reconstruction

C.3 Matched response ablations

Within each of the 512-session and 64-session ablations, all receiver conditions use identical audio, source/action probabilities, state maps, key panels, and payload schedules. Tables 10 and 18 report the respective results. Only the transition used in Equation (15) changes. Identity uses $W = I$; native fitting uses 192 native pairs; full pooled fitting uses 576 views. The equal-budget pooled arms use 64 native, 64 DAC, and 64 SNAC pairs, in three rotations over the same 192 source clips. Source probabilities and action laws remain fixed across arms.

Each receiver is evaluated at fixed endpoints after two and three observations, giving eight-second and twelve-second joint recovery, respectively. False positives count acceptance of the queried key on unmarked sessions at each endpoint.

C.4 Key and owner protocols

The main 512-session music benchmark and the supplementary 64-session benchmark use the owner key K_0. Each panel contains K_0 and $99{,}999$ decoys generated by domain-separated HMAC-SHA256 expansion under K_0. The panel remains fixed across receiver substitutions. These evaluations report empirical fixed-panel acceptance and recovery. The exchangeable-key theorem in Appendix A.6 applies to the stated randomized-key experiment; its 10^{-5} bound is not assigned to these owner-derived panels.

The new-owner experiment uses a frozen pooled response dictionary and evaluates eight new owner keys with new nonces and audio while holding the 64-prompt set fixed. Each owner carries all sixteen payloads, giving 128 marked sessions and 128 unmarked controls. Owners are crossed with payloads over the fixed prompt set, without a full owner–payload–prompt crossing. Each session is scored at twelve seconds. Single-key and any-of-eight false positives are reported separately.

The speech panels use four new 32-byte owner keys from one pseudorandom stream and 99,996 reference keys from a separately seeded stream, forming a 100,000-key panel. Model-selection and neural-codec panels use different seeds. Their owner–payload cells are set before scoring. The decision has one look after six observations.

The two-look music analysis in Appendix E.5 adds a third observation when the first two do not produce acceptance. This sequential rule is distinct from the fixed eight- and twelve-second ablation endpoints. Scoring can reject through a 1,000-key prefix: a reference already at or above the candidate prevents that candidate from being uniquely first in every extension of the panel. For a multiple-look decision, this rejection condition must hold at every allowed look. Potential acceptances are evaluated against the complete panel.

C.5 Owner-absent evaluation

A 100,000-key panel excludes the generating key on every marked input. For each channel, 128 marked and 128 unmarked sessions are scored at twelve seconds. One indexed candidate supplies the single-query result. The first 256 indexed keys supply 32,768 queries for each marked or unmarked channel cell, or 393,216 queries across both input types and six channels.

The endpoint is a unique rank-one event in the full panel. The single-query result is contained in the 256-query result. All queried labels on an input use the same ranking, and transformed views share their source audio. Counts are therefore nested and dependent. Equation (34) gives a conditional key-

symmetry expectation for this design, not an independence-based significance test or a per-fixed-key guarantee.

C.6 Baseline detection and message aggregation

Main comparison. The 512-session comparison uses the independently calibrated operating points in Appendix C.1. HPT uses $h = 0$, $\gamma = 0.25$, and $\delta = 1$, with active heads 0–3 and frozen Leiden maps. The native codec is not fine-tuned. Appendix D.2 reports conservative ranks and tied-maximum counts for its generating key.

Supplementary comparison. The Aligned-IS and AudioSeal details below describe the 64-session comparison. Their nominal thresholds are separate from the calibrated operating points of the main benchmark.

Aligned-IS. The MusicGen implementation uses twenty clusters per head and context length one on all four heads. Euclidean KMeans with ten initializations is applied to the first 2,048 decoder input embeddings, followed by capacity-balanced Hungarian assignment with capacity $\lceil 2048/20 \rceil$. Top-250 truncation precedes watermark reweighting. Insertion excludes repeated previous-token contexts within an observation and head; the public detector counts the 199 observed transitions of each head. Session decisions pool hit and trial counts across all three observations, preserving their boundaries. The implementation's nominal one-sided binomial rule uses probability $1/20$ and threshold 10^{-5}.

AudioSeal. We use the official base sixteen-bit generator and detector without retraining. The watermark is computed at 16 kHz, resampled to 32 kHz, and added at unit strength. Presence is determined by the mean of the three observation-level mean watermark probabilities, thresholded at 0.5. Bit probabilities are clipped to $[10^{-7}, 1-10^{-7}]$, converted to logits, averaged across observations, and thresholded at zero.

The sixteen application messages are represented by balanced codewords. For message m and bit position $j \in \{0, \ldots, 15\}$,

$$c_{m,j} = \mathrm{parity}\big[(\lfloor m/2 \rfloor + 1) \,\&\, j\big] \oplus (m \bmod 2). \tag{43}$$

Each word has eight ones and the minimum Hamming distance is eight. The application decoder chooses the unique nearest codeword; ties abstain. Joint recovery requires both a positive presence decision and the correct application message.

C.7 Speech selection and response comparisons

Selection panel. Direct-response candidates vary the state count over $\{16, 32, 64, 128\}$ and smoothing over $\{1, 4, 16\}$. All predict the same 48 raw-token action targets for heads 0 and 1 and $J \in \{8, 16\}$, each with weight $1/48$. The deployed speech program uses the sixteen $J = 8$ targets; selection evaluates the full dictionary. Fitting uses 72 unmarked observations from eight text-source groups. Selection uses three observations for each of sixteen other groups. A third set of sixteen groups, again with three observations each, is reserved for held-out risk evaluation. All observations, views, and rows of a source remain together.

Table 9 separates the three unmarked stages. The empirical loss averages squared target errors within an observation, observations within a source group, and then source groups equally. Thus the source group, rather than a row or transformed view, is the resampling unit. Ties in selection prefer fewer states and then larger smoothing. The selected 32-state, $\alpha = 1$ receiver is fixed before risk assessment. Confidence intervals use 10,000 paired percentile bootstrap resamples of the sixteen held-out groups, applying each resample to both selected and neutral losses.

Marked evaluation crosses four new owners with sixteen payloads. The 64 text sources are distinct English passages of 60–75 words. Each has six marked and six unmarked observations. The channels are native rendering, FLAC, Opus at 96 kbit/s, and MP3 at 128 kbit/s. MP3 is excluded from fitting and selection. Source-emission and posterior controls are evaluated on the same marked and unmarked sessions.

Table 9: Unmarked speech selection protocol. Source groups are disjoint across stages; all rows and views belonging to a source remain together. These counts describe the 32-state selection evaluation.

Stage	Source groups	Observations	Role
Fitting	8	72	Estimate each candidate's responses
Selection	16	48	Choose state count and smoothing
Held-out assessment	16	48	Evaluate the frozen selected predictor

Neural-codec panel. The speech neural-codec evaluation compares native Mimi, DAC, SNAC 32 kHz, and both compositions. Every fitted arm uses 72 fitting and 48 selection observations with the same source-group allocation. The pooled fit distributes its fitting examples as evenly as possible across the five channels; leave-one-out fits allocate eighteen examples to each included channel. The same unmarked criterion selects states and smoothing. SNAC 24 kHz is excluded from fitting and selection and used only for evaluation. The unmarked criterion selects sixteen states and unit smoothing for the pooled neural-codec receiver. Each arm uses its own fitted response dictionary.

Fitting-budget panel. The data-budget study fixes 64 states and smoothing sixteen, using nested subsets of a 576-pair pool from 64 sources. Budgets are 32, 64, 128, 256, and 576 pairs. Three subsampling seeds are used below the full budget. The smallest subset includes 32 sources and the others include 64. A separate 48-observation selection set supplies the prediction diagnostic; marked evaluation uses the direct-response evaluation sessions. Because this sweep fixes a different estimator and prediction-evaluation split, its loss values are interpreted within the sweep rather than compared numerically to Table 15.

C.8 Metrics and additional panels

A joint success is correct owner acceptance with exact payload recovery. A supplied-key payload success is the correct message maximizer, irrespective of owner acceptance. An incorrect marked acceptance returns a wrong owner or payload. Null attribution is acceptance on an unmarked input for the queried key (single-key) or for at least one key in the tested owner family (any-owner), as specified.

AUC compares marked and unmarked continuous scores, assigning half credit to ties. The supplementary music comparison uses 64 sessions of each kind and the new-owner study uses 128. CHORD's common-comparison score is the twelve-second owner statistic $A(k)$, Aligned-IS uses its session presence statistic, and AudioSeal uses its mean watermark probability.

The source-disjoint music evaluation contains 640 marked sessions and their controls. An eight-owner/new-prompt study crosses eight keys with eight new prompts and assigns two Latin-balanced payloads to each owner–prompt cell, giving 128 sessions. Every owner and prompt encounters all sixteen payloads. FMA-small supplies 4,096 source-unique recorded tracks (Defferrard et al., 2017).

Further music codec transfer uses sixteen sessions under SpeechTokenizer, Mimi, and NeMo-FSQ, with 32 controls per codec. The signal-edit panel has 32 marked cases and sixteen controls per condition. The speech prefix study fixes the selected direct receiver and scores prefixes of the first observation rather than the full six-observation session.

D Detailed Main-Benchmark Results

This section presents detailed results for the 512-session music benchmark, new-owner reuse, and the speech response comparisons. The supplementary music panels in Appendix E use their own stated session counts. Comparisons within a panel use paired views or receiver substitutions; totals from different panels are not combined.

D.1 Response fitting and attribution

The matched music study fixes the sender, audio, source action probabilities, state map, key panel, and decision rule. The count-matched pooled arms rotate 64 pairs from each of the three fitting

channels. Each receiver is evaluated at fixed eight- and twelve-second endpoints. The tests compare receivers on matched sessions using two-sided exact McNemar tests with Holm correction across all twelve channel–length tests, including the saturated Native and DAC comparisons. The outcome table separates joint success from correct supplied-key payloads whose owner is rejected. It is an analysis of the evaluated decisions, not an estimator of the owner/payload contrast norms.

Table 10: Response learning with the sender fixed. Joint successes at eight and twelve seconds, out of 512 sessions. Native and DAC are at $512/512$ throughout. p is the exact McNemar p-value against identity at twelve seconds, Holm-corrected over twelve channel–length tests. The twelve-second summary appears in Table 1.

Response	Pairs	SNAC	DAC → SNAC	SNAC → DAC	NeMo-FSQ
Identity	—	366 → 475	239 → 385	234 → 379	498 → 508
Native	192	496 → 512	435 → 503	437 → 506	512 → 512
Pooled	576	496 → 512	435 → 501	434 → 505	512 → 512
Pooled (rot.)	192	493–497 → 512	435–436 → 500–502	433–436 → 504–506	512 → 512
p, identity vs. pooled (12 s)		1×10^{-10}	6×10^{-33}	2×10^{-37}	0.63

Table 11: Outcomes at twelve seconds, out of 512 sessions per channel–response pair. Native and DAC, at 512 joint successes under both responses, are omitted. No unmarked session is accepted for the queried key in any cell.

Channel	Response	Joint	Payload correct, owner rejected	Payload wrong
SNAC	Identity	475	37	0
	Pooled	512	0	0
DAC → SNAC	Identity	385	126	1
	Pooled	501	11	0
SNAC → DAC	Identity	379	132	1
	Pooled	505	7	0
NeMo-FSQ	Identity	508	4	0
	Pooled	512	0	0

Across the four rows in Table 11, the identity receiver has 301 joint failures: 299 retain the correct supplied-key payload and two have an incorrect payload. The pooled receiver has 18 joint failures, all retaining the correct payload. Channel views share source sessions, so these are session–channel counts, not 301 independent recordings. On DAC-to-SNAC specifically, the reported paired outcomes are 117 pooled-only successes and one identity-only success, yielding the 116-session net improvement in Table 1.

D.2 HPT GENERATING-KEY RANKS

For a queried key, define the conservative rank as one plus the number of reference keys with score at least as large. Table 12 reports this statistic for the generating HPT key on the 512 native twelve-second sessions. The key is never uniquely or jointly first, and never enters the top 1,000. Its median rank is 5,571. The 0.996632 AUC uses *negative conservative rank* to distinguish marked from unmarked sessions; it measures marked/null discrimination, whereas unique-first acceptance measures key identification. The simultaneously high rank-based AUC and absence of a unique first-place generating key show why these endpoints must be reported separately. They do not establish the cause of HPT's rank distribution or its behavior under other configurations.

Table 12: HPT on native twelve-second music sessions. Rank statistics use the 512 marked sessions; the rank-based AUC compares marked and unmarked sessions.

Diagnostic	Result
Calibrated presence TP	$464/512$
Generating key uniquely first	$0/512$
Generating key tied for first	$0/512$
Generating key in top 1,000	$0/512$
Conservative rank: minimum	1,347
Conservative rank: median	5,571
Conservative rank: maximum	15,688
Marked/null AUC using negative conservative rank	0.996632

D.3 New-owner reuse

Eight new owner keys each carry all sixteen payloads while holding the 64-prompt set fixed. The fitted response dictionary is frozen across owners. Table 13 summarizes joint recovery, supplied-key decoding, and both null-attribution endpoints. Because entries are indexed by action and observed state, reuse is expected by construction; this panel checks it empirically without key-specific fitting.

Table 13: New-owner reuse with a frozen music receiver. Each channel has 128 marked and 128 unmarked sessions, scored at twelve seconds. Single-key FP counts acceptances for the queried key; Any-of-eight FP counts sessions accepted by at least one of the eight keys.

Channel	Joint	Supplied-key payload	Single-key FP	Any-of-eight FP
Native	128	128	0	0
DAC	128	128	0	0
SNAC	128	128	0	0
DAC $\rightarrow$ SNAC	127	128	0	0
SNAC $\rightarrow$ DAC	127	128	0	0
NeMo-FSQ	128	128	0	0

D.4 Speech response comparison

The neural-codec panel in Table 2 uses a 16-state pooled receiver with unit smoothing; the native/FLAC/Opus selection evaluation uses a 32-state receiver. Table 14 reports the paired recovery outcomes: each row partitions the same 64 sessions into successes for both receivers, source emission only, pooled response only, and neither. These pairs compare source emission with pooled response learning; Table 2 also reports channel-specific and leave-one-out totals.

Table 14: Paired joint-recovery outcomes in the 64-session neural-codec speech panel. Counts partition the same sessions and reproduce the source-emission and pooled totals in Table 2. Each session totals 96 seconds. SNAC 24 kHz is excluded from fitting and selection. p is the uncorrected two-sided exact McNemar p-value on the discordant sessions.

Channel	Both	Source only	Pooled only	Neither	p
SNAC	54	0	7	3	0.016
DAC $\rightarrow$ SNAC	53	0	4	7	0.125
SNAC $\rightarrow$ DAC	53	0	4	7	0.125
SNAC 24 kHz	50	1	7	6	0.070

All seven SNAC gains and the four gains on each composition occur without a regression on the same channel. The held-out operating point has seven gains and one regression, for six additional successes overall. The paired outcomes complement aggregate recovery by separating gains from regressions.

Held-out prediction risk. For the 32-state selection panel, the held-out risk reduction is 0.0307 (95% interval $[0.0257, 0.0353]$). Table 15 normalizes this reduction by the neutral predictor's risk,

Table 15: Held-out prediction risk after unmarked receiver selection. The 32-state receiver predicts 48 fixed action targets. Intervals use paired bootstrap resampling of sixteen held-out source groups.

Quantity	Estimate	95% interval
Selected receiver risk	0.914	[0.885, 0.941]
Neutral response risk	0.944	[0.918, 0.970]
Risk reduction	0.031	[0.026, 0.035]
Relative risk reduction	3.3%	[2.7%, 3.8%]

giving 3.3%. This is a prediction-risk improvement on fixed queries. By Appendix A.2, it estimates a lower bound on the weighted fraction of source query χ^2 surviving into the 32-state representation, not that fraction itself.

D.5 Removal, synchronization, and observation-model changes

The frozen receiver is evaluated on the nine conditions in Table 16, including two noise levels and two DCCRN input conditions. Search re-evaluates the same audio under the reported number of candidate hypotheses, keeping the fitted receiver fixed. Offset and rate hypotheses address temporal correspondence; polarity inversion is a sign change, not a temporal shift. The search-result columns are empirical outcomes. Under the exchangeability assumptions of Section 4.3, L fixed looks would give the per-query union bound $L/100{,}000$. The bound column is this conditional reference calculation, not a certified false-positive rate for the fixed-key experiment.

Table 16: Removal attempts, 64 marked and 64 unmarked sessions per condition, twelve seconds, same frozen receiver. L is the number of reported looks; the bound column assumes exchangeability and fixed looks. It is not an empirical fixed-key guarantee. ΔCLAP is the marked text–audio similarity relative to the same audio before the transformation. No condition produces a false acceptance or an accepted wrong payload.

Transformation	Fixed	With search	L	Bound	ΔCLAP
Noise, 20 dB SNR	63	—	1	10^{-5}	-0.016
Noise, 30 dB SNR	64	64	1	10^{-5}	$+0.000$
Polarity inversion	0	**64**	2	2×10^{-5}	$+0.000$
Delay, 10 ms	0	**64**	5	5×10^{-5}	$+0.000$
Speed, $1.1\times$	0	**64**	15	1.5×10^{-4}	-0.065
Time stretch, $0.9\times$	0	1	15	1.5×10^{-4}	$+0.026$
Pitch, $+1$ semitone	0	1	15	1.5×10^{-4}	$+0.028$
DCCRN enhancer, clean input	0	0	1	10^{-5}	-0.153
DCCRN enhancer, 30 dB input	0	0	1	10^{-5}	-0.166

Time stretching and pitch shifting each produce one joint success with the reported searches. Both DCCRN conditions use $L = 1$ and produce zero joint successes; no multi-look DCCRN recovery is evaluated here. Recovery after delay/rate search diagnoses sensitivity to correspondence, while the remaining failures do not by themselves establish erasure of source evidence. The CLAP differences compare each transformed waveform with its pre-transformation counterpart; the marked–unmarked quality measurements below use a different pairing. Table 29 reports additional signal edits, and Appendix E.9 reports the crop-and-pad evaluation.

D.6 Main-panel quality measurements

The CLAP results compare methods on the same 512-session evaluation. Intervals containing zero indicate no detectable paired difference, rather than equivalence. Under the tested metric, the intervals for HPT and AudioSeal exclude zero. Table 26 reports the 64-session supplementary quality evaluation.

Table 17: Paired text–audio similarity relative to unmarked generations in the 512-session evaluation, measured with frozen CLAP (Wu et al., 2023). Speech word and character error rates use 64 passages. Intervals are paired-bootstrap 95% confidence intervals.

Method	Marked	Marked − unmarked	95% CI
CHORD	0.5825	-0.0013	$[-0.0056,\ 0.0029]$
Aligned-IS	0.5851	$+0.0013$	$[-0.0030,\ 0.0055]$
HPT	0.5727	-0.0111	$[-0.0165,\ -0.0059]$
AudioSeal	0.5809	-0.0029	$[-0.0034,\ -0.0024]$
CHORD, speech WER	1.50%	-0.15 pt	$[-0.39,\ 0.09]$
CHORD, speech CER	0.85%	-0.06 pt	$[-0.21,\ 0.08]$

E Supplementary Evaluation Panels

This appendix reports supplementary music, speech, and signal-transformation evaluations. Session counts and receiver configurations are specified for each panel. The 64-session response ablation and configuration-sensitivity evaluation use different audio sets; results are reported within each panel.

E.1 Supplementary 64-session music response ablation

Table 18 gives every response condition from the matched ablation. At twelve seconds the identity receiver recovers all supplied-key payloads. The learned responses improve joint recovery through the owner decision. No false acceptance of the queried key occurs among the 64 controls for any response, channel, or length.

Table 18: **Complete response ablation.** Joint successes out of 64, written as 8 s / 12 s. Rotations vary the channel assignment of the equal-budget pooled fit.

Response / pairs	Native	DAC	SNAC	DAC → SNAC	SNAC → DAC	NeMo-FSQ
Identity	64 / 64	64 / 64	50 / 61	27 / 52	37 / 54	64 / 64
Native / 192	64 / 64	64 / 64	61 / 64	57 / 64	58 / 64	64 / 64
Pooled / 576	64 / 64	64 / 64	61 / 64	58 / 64	58 / 64	64 / 64
Pooled / 192, rotation 1	64 / 64	64 / 64	61 / 64	58 / 64	57 / 64	64 / 64
Pooled / 192, rotation 2	64 / 64	64 / 64	61 / 64	57 / 64	58 / 64	64 / 64
Pooled / 192, rotation 3	64 / 64	64 / 64	61 / 64	57 / 64	57 / 64	64 / 64

E.2 Music configuration sensitivity

The MusicGen configuration-sensitivity evaluation compares twelve receiver configurations using unmarked prediction loss. The criterion selects 64 states and smoothing $\alpha = 16$. All twelve candidates have identical twelve-second recovery counts (Table 19).

The main music estimator uses marginal prior mass one. Candidate-wise losses and margins are not reported; the observed recovery plateau does not establish that lower prediction loss yields higher recovery.

E.3 Baseline outputs on the supplementary 64-session panel

Table 20 separates presence decisions from message recovery. On native audio, AudioSeal's application codebook corrects one nonexact physical string. Under the more disruptive codecs, the low message recovery accompanies a decrease in bit accuracy, in addition to the presence decision's loss of power.

Table 19: MusicGen configuration sensitivity. All twelve candidates have the listed twelve-second recovery counts; unmarked loss selects 64 states and $\alpha = 16$.

Channel	Joint recovery for each candidate
Native	$64/64$
DAC	$64/64$
SNAC	$64/64$
NeMo	$64/64$
DAC $\rightarrow$ SNAC	$63/64$
SNAC $\rightarrow$ DAC	$63/64$

Table 20: **Presence and message outputs.** TP, FP, and message counts are out of 64 sessions. CHORD has zero single-key FP in every row. Application recovery is shown before the AudioSeal presence decision.

	Aligned-IS	AudioSeal			
Channel	TP / FP	TP / FP	Bit accuracy	Exact 16 bits	Application msg.
Native	64 / 0	64 / 0	99.9%	63	64
DAC	64 / 0	62 / 0	95.4%	36	63
SNAC	18 / 0	0 / 0	50.6%	0	1
DAC $\rightarrow$ SNAC	5 / 0	0 / 0	52.1%	0	3
SNAC $\rightarrow$ DAC	7 / 1	0 / 0	54.0%	0	2
NeMo-FSQ	49 / 1	0 / 0	52.9%	0	0

E.4 Owner-absent queries

Table 21 reports the two observed events and their locations. The single indexed query is included in the first-256 set, so the two counts are nested. There are 1,536 processed audio inputs across the six channels and two audio types, with 256 queried labels per input. Under exchangeability and an almost surely unique maximum, the reference expected count is $393{,}216/100{,}000 = 3.93216$. This conditional expectation does not treat the dependent queries as independent trials.

Table 21: **Unique-top events with the generating key excluded.** Each channel has 128 marked (M) and 128 unmarked (N) sessions. First 256 contains 32,768 queries of each type per channel.

Channel	Single indexed query, M / N	First 256, M / N
Native	0 / 0	1 / 0
DAC	0 / 0	0 / 0
SNAC	0 / 0	0 / 0
DAC $\rightarrow$ SNAC	0 / 0	0 / 0
SNAC $\rightarrow$ DAC	0 / 0	0 / 0
NeMo-FSQ	0 / 0	0 / 1

E.5 Additional music reuse and observation length

The evaluations in table 22 use the fixed music receiver and distinct source panels. The source-disjoint marked miss retains a correct supplied-key payload. The two channel rows in the eight-owner/new-prompt study share their 128 source sessions.

In the two-look analysis of 64 music sessions under six channels, 369 of the 384 marked channel views are accepted after eight seconds; fifteen use the third observation. All accepted owners and payloads are correct. Mean consumption is 8.2 seconds. Native, DAC, and NeMo-FSQ require no third observation; SNAC requires it in three sessions and each composition in six. These durations sum the consumed, separately generated observations. The adaptive two-look rule differs from the fixed-length recovery endpoints.

Table 22: **Additional reuse evaluations.** These panels are separate from the new-owner experiment in table 13.

Evaluation	Endpoint	Result
Source-disjoint marked	Joint owner–payload	639/640
Source-disjoint marked	Supplied-key payload	640/640
Source-disjoint marked	Incorrect accepted message	0/640
Source-disjoint controls	False owner attribution	1/640
Eight-owner/new-prompt panel, native	Joint owner–payload	128/128
Eight-owner/new-prompt panel, SNAC → DAC	Joint owner–payload	128/128
Eight-owner/new-prompt controls	Any-owner attribution	0/64
FMA-small	False owner attribution	0/4,096

E.6 Speech model selection and response learning

Table 15 reports the held-out assessment following unmarked selection. The relative reduction in prediction risk is $1 - \widehat{L}(f)/\widehat{L}(\mathbf{1})$. The selected receiver and the projected controls are compared in table 23. All listed controls recover the supplied-key payload in every session; no unmarked session is assigned to any of the four owners. The selected receiver ties the same-state source-emission control on native/FLAC/Opus and has one fewer MP3 joint success. Prediction risk and recovery are distinct endpoints.

Table 23: **Speech response controls.** Joint successes out of 64 using the same 96-second sessions. B is state count and α is smoothing. Native includes waveform rendering and Mimi re-encoding.

Receiver	B / α	Native	FLAC	Opus	MP3
Selected direct response	32 / 1	64	64	64	61
Source emission	32 / 1	64	64	64	62
Source emission	64 / 16	64	64	64	62
Hard source posterior	64 / 16	64	64	64	62
Soft source posterior	64 / 16	64	64	64	61
Raw-token source response	2,048 / 16	63	62	62	48

Table 14 gives the paired neural-codec outcomes. Each direct-response arm recovers every supplied-key payload; source emission recovers 62/64 on DAC-to-SNAC and all payloads elsewhere. Any-of-four null attribution is zero throughout.

Speech observation budget. Table 24 evaluates the selected speech receiver on prefixes of the first observation. All accepted marked cases have the correct owner and payload, and all unmarked attribution counts are zero. A secondary rule that permits all four prefix looks gives 5/64 native and FLAC, 4/64 Opus, and 2/64 MP3 joint successes. The main 96-second sessions accumulate separate observations and are not single short excerpts. The four-look and fixed-prefix results also use different decision rules.

Table 24: **Speech inference from a single-observation prefix.** Each entry is joint / supplied-key payload successes, both out of 64.

Duration	Native	FLAC	Opus	MP3
2 s	1 / 28	1 / 28	1 / 29	1 / 22
4 s	0 / 33	0 / 33	0 / 31	0 / 28
8 s	1 / 46	1 / 46	1 / 46	0 / 41
12 s	4 / 55	4 / 55	3 / 54	2 / 53

E.7 Unmarked fitting budget

The budget comparison fixes the representation and estimator while varying the number of fitting pairs. With 32 pairs, the direct receiver already obtains at least 61 joint successes per condition. The

reported prediction loss decreases over the budget sweep, and the full-budget receiver recovers every session in the tested channels (table 25). The sweep uses a different fixed configuration from the 32-state main speech receiver. Supplied-key payload recovery is 64/64 throughout, and all fits have zero any-owner attribution on 64 controls per channel.

Table 25: **Data efficiency of direct response learning.** The receiver has 64 states and smoothing sixteen; joint counts are out of 64. Ranges summarize three subsampling seeds. Contrast loss uses a separate unmarked selection set.

Pairs	Sources	Native	FLAC	Opus	MP3	Contrast loss
32	32	63–64	63–64	63–64	61–63	0.927–0.929
64	64	63–64	63–64	63–64	61–63	0.924–0.925
128	64	64	64	64	62–63	0.923
256	64	64	64	64	62–63	0.921–0.922
576	64	64	64	64	64	0.921

E.8 Audio quality

For the common music comparison, frozen `laion/clap-htsat-unfused` measures text–audio similarity (Wu et al., 2023). Three observations are concatenated, saved and read as PCM16, and resampled to 48 kHz. A shared per-session seed gives the same ten-second crop across methods and marked/unmarked outputs. Cosine similarity between normalized text and audio embeddings supplies the score. Table 26 reports differences paired by prompt for this supplementary panel.

Table 26: **Music text–audio similarity.** The supplementary comparison has 64 paired sessions and unmarked mean 0.565. The new-owner evaluation uses 128 paired differences.

Method / evaluation	Mean score	Mean Δ	Median Δ	Minimum Δ
Aligned-IS	0.567	+0.002	+0.013	-0.119
AudioSeal	0.559	-0.005	-0.004	-0.018
CHORD, common comparison	0.558	-0.006	-0.006	-0.114
CHORD, new owners	—	+0.009	+0.006	—

Speech intelligibility uses `whisper-small.en` with FP32 eager greedy decoding. Error rates are computed on the full saved passages and averaged equally over source groups; watermark inference uses the first sixteen seconds of each observation. Each condition has 384 marked and 384 unmarked waveforms from 64 sources.

Table 27: **Speech word error rate (%).** The direct-selection and neural-codec panels are separate evaluations.

Evaluation / channel	Marked	Unmarked
Direct selection / native	1.50	1.65
Neural-codec study / native	1.70	1.65
Neural-codec study / DAC	1.61	1.82
Neural-codec study / SNAC	2.16	2.16
Neural-codec study / DAC $\rightarrow$ SNAC	2.21	2.31
Neural-codec study / SNAC $\rightarrow$ DAC	2.12	2.23
Neural-codec study / SNAC 24 kHz	2.46	2.75

E.9 Further codecs and signal transformations

The additional music codecs in table 28 are excluded from response fitting. All supplied-key payloads are recovered, with lower owner acceptance on SpeechTokenizer and Mimi (Zhang et al., 2024; Défossez et al., 2024). All unaccepted marked cases abstain.

Table 28: **Additional music codec transfer.** Sixteen marked sessions and 32 controls per codec.

Codec	Payload	Owner + payload	Incorrect accept.	Null attribution
SpeechTokenizer	16/16	9/16	0/16	0/32
Mimi	16/16	9/16	0/16	0/32
NeMo-FSQ	16/16	16/16	0/16	0/32

Table 29: **Music transformations.** Every condition uses 32 marked cases and sixteen controls. The same fixed-frame receiver is used throughout.

Transformation	Owner + payload	Incorrect accept.	Null attribution
MP3, 64 kbit/s	32/32	0/32	0/16
AAC, 64 kbit/s	32/32	0/32	0/16
Opus, 32 kbit/s	32/32	0/32	0/16
Noise, 20 dB SNR	32/32	0/32	0/16
Low-pass, 3 kHz	32/32	0/32	0/16
Resampling, 16 kHz	32/32	0/32	0/16
Echo, 300 ms	32/32	0/32	0/16
Polarity inversion	0/32	0/32	0/16
Speed, $1.1\times$	0/32	0/32	0/16
Pitch, +1 semitone	0/32	0/32	0/16
Time stretch, $0.9\times$	0/32	1/32	0/16

Crop-and-pad evaluation. On the tested `crop50_zeropad` operation, joint recovery is $32/32$ with zero false attributions among sixteen unmarked controls. This 50% crop-and-pad result does not establish arbitrary-crop robustness.