

Learning, expectation and musical pleasure: Supplementary Methods and amendment record

Amaç Erdem | amace@bu.edu | ORCID 0009-0000-1711-0036

Abstract. This supplement specifies the integrated human-study and MI-evaluation programme dated 12 October 2026. It separates current design commitments from pilot-set quantities, optional future assays and superseded settings. Detailed measurement, model, inference and infrastructure contracts support a compact main narrative. The prior 35 figure sources are retained through 15 combined archival plates and in the editable package; their earlier numerical settings are explicitly identified as historical. The independent LLM-analysis benchmark remains a separate future experiment. No participant, neural, benchmark or operational-validation results are reported.

Current specification: S1–S10. Separate future studies: S11. Historical figure record: Appendix A. The dated main protocol and this amendment take precedence over the earlier papers.

S1. Objectives, decisions and amendment ledger

Aim	Measurement	Matched comparison	Decision
Human learning	Pre-outcome chance of A; assigned and realized exposure	Changed versus stable continuation schedule	Changed-minus-stable change in $q(A)$; report effect and uncertainty
Pleasure relationships	Expected enjoyment, experienced pleasure, continuous trace, recognition	Within-person/motif association with acoustic and exposure controls	Prespecified secondary relations; no neural or unrestricted latent-state claim
MI contribution	Within-excerpt trajectory and between-excerpt trace mean	Updated personal MI versus equally updated acoustic learner on identical targets	Joint meaningful gain requires $\Delta r > 0.10$ for each scale with confirmatory evidence
Personalization	Same target and representation	Population versus calibrated versus earlier-feedback-updated state	Separate history and updating increments; secondary family
Mechanism diagnosis	Matched model outputs and target-relevant reports	Leave-one-module/group-out, temporal controls and shared-input audits	Locate useful/weak operationalizations; revise only after primary freeze
Measurement qualification	Gesture practice, timing, interruptions, burden, passive/report blocks	Known control signals and counterbalanced task conditions	Pilot-exit qualification or redesign before confirmation

The programme has two coequal scientific aims, not one scientific aim plus a promotional evaluation. Listener updating and model updating are separate processes. MI is a literature-anchored computational proposal: source motivation, input selection, equation, temporal support and empirical target are recorded separately. The nine-mechanism source audit found citation metadata and MI-defined operational choices in all nine modules. This is neither a claim of 100% copied formulas nor a quantified claim of field-wide novelty.

Substantive amendments from 11 October. (1) Exactly 20 previously recognized works and at least 16 usable ratings are replaced by three-source calibration, with dose/minimum fixed at pilot exit. (2) Desktop/laptop eligibility is replaced by qualified mobile collection. (3) The three-alternative probability-allocation task and 1:1 U/L assignment leave the core: a two-alternative updating assay supports one-gesture probability reports; a three-alternative uncertainty assay remains optional. (4) Two trajectory/mean correlation increments above 0.10 replace the one-point absolute-error improvement as the primary practical model target. (5) The trace mean becomes the between-excerpt primary target; immediate overall rating remains a distinct secondary outcome. (6) Three earlier primary claims are replaced by two confirmatory families: human updating and joint MI gain. (7) The 55–75-minute duration is a historical estimate, not a promise for mobile collection. (8) Human measurement/inference methods are integrated; the LLM benchmark is independent.

Retained starting values include the 1,000-registration cap with an initial 60-registration pilot, 24 transfer excerpts of 25 seconds, four six-slot policies, four secondary performance pairs, 5-Hz sampling and 5,000 bootstrap replicates. These are proposed design settings. The existing 16-training/32-probe learning schedule supplies the first pilot candidate, with switch after probe 16 and before/after windows 9–16/25–32. A lower-burden schedule needs recovery and feasibility justification, then an explicit amendment.

Decision status. Fixed scientific commitments are the two aims, honest information boundaries, matched comparators, separate outcome scales and $\Delta r > 0.10$ target. Pilot-set quantities include calibration dose, response lag, task length, target-based masks, transformations, final sample and power alternative. Optional studies include full uncertainty distributions, value-learning extensions, independent neural observations and the LLM benchmark. None can be silently substituted for an unsuccessful primary endpoint.

Source preservation. The preceding three PDFs, their exact TeX/reference sources, 35 original SVGs, 15 combined SVG/PDF plates and the previous technical supplement remain in the editable archive. Appendix A reproduces the 15 combined plates. A machine-readable section/figure/reference crosswalk records retained, amended, deferred and archived material. Historical diagrams preserve their content; captions identify their non-current settings rather than presenting them as current requirements.

S2. Recruitment, catalogue and personal calibration

Population and consent. Adults aged 18 or older enter through language/device/headphone checks and versioned consent. Registration, eligibility and study participation are different events. The cap counts unique eligible registrations including pilot registrations; incomplete cases remain in the flow. Verification establishes account control rather than uniqueness of a person. Record invitations, screening, consent, calibration completion, task readiness, started/completed blocks and endpoint-specific inclusion. State compensation and withdrawal procedures independently of agreement with MI. Translation requires cognitive interviewing and counterbalanced qualification before pooling.

Catalogue identity. Keep composition family, movement, arrangement, performance, master recording, excerpt and exposure identifiers. Different encodings do not create independent music; a content hash cannot establish composition identity. Use exact rights-cleared recordings or newly rendered controlled motifs, with documented permissions and stable audio preprocessing[1]. Standardized fades, channel layout and fixed loudness normalization preserve within-excerpt dynamics. Reject clipping, unresolved metadata and unmatched passage boundaries. Retain loudness, roughness, tempo, onset density, tonal stability and instrumentation controls. Arrangement changes fall outside the paired-performance inference unless separately designed.

Three routes. Route K samples previously recognized pieces, including liked/disliked anchors. Route P samples high and low predicted-preference candidates from the model state available before their ratings. Route N samples listener-declared unfamiliar pieces. Routes describe selection; familiarity and predicted preference are separately coded. A recommended unfamiliar item can satisfy both attributes but receives one exposure identity and a declared route assignment. Item selection must preserve score range and acoustic coverage; selecting only predicted favourites would answer a restricted policy question.

The pilot evaluates 12-, 18- and 24-reference designs with roughly balanced route budgets, adjusting overlap by a deterministic predeclared allocation rule. Begin with a common seed subset; initial population priors may rank candidates before personal responses exist. Subsequent recommendations use only completed seed/reference responses. Save the selecting model and its information cutoff for each item. A 12-second preview is the initial candidate duration, subject to coverage/burden qualification. It counts as exposure even if recognition fails. Do not infer streaming history or autobiographical memory, and do not require an account with a music service.

Personal state. Record prior recognition, graded familiarity, performance recognition and pleasure separately, with an unable-to-rate option. No-response is not zero pleasure. Save interrupted calibration and shrink weakly supported parameters toward the pilot-trained population prior. The pilot-exit rule specifies usable-reference counts, route coverage and readiness; it supersedes the older 16-of-20 rule. Reference-dose analysis on the same held-out targets remains secondary, using nested prespecified subset orderings and common fitting resources.

Transfer selection. Before ranking, randomly partition an eligible candidate pool into four policy candidate sets; rotate membership across people. Policies are personal MI, matched acoustic, matched random and matched swapped-profile MI. Personal and acoustic selectors use the same admissible history and ranking budget. Match the random policy on declared broad familiarity, instrumentation and duration; do not match on the MI score being tested. Match swapped profiles on training/catalogue coverage without revealing identities. Unfillable quotas prevent release until a prespecified fallback or amendment applies. Save inclusion probabilities, seed and hidden ranks.

Interleave six slots per policy in balanced blocks, with no policy label in the interface or filenames. The 24-excerpt released mixture defines the primary population of test items. Every predictor scores every delivered observation, irrespective of selector. Thus accuracy is distinct from policy value. Generalization to an unrestricted catalogue requires a separately sampled random audit or known-probability weighting with support; this protocol does not assert it from selected excerpts alone. Report policy-specific sensitivities.

Four secondary paired-performance comparisons hold composition passage and broad instrumentation constant while comparing two previously unheard performances beyond the reference. Counterbalance performance order; obtain both overall ratings and a preference response with an indifferent option. A preview or repeat is never treated as first exposure. Store actual delivery history even when a participant reports outside-study familiarity.

S3. Human learning, expectation and pleasure

Low-burden core. Demonstrate continuation alternatives A and B with audible examples; notation knowledge is unnecessary. A single vertical response indicates the probability of A, with B defined as its complement. Practice checks interpretation through unambiguous probability examples; record failures before random assignment. Probability, confidence, expected enjoyment, surprise and experienced pleasure are different variables. Only one is requested per screen. Confidence is optional secondary, not an obligatory extra step at each core probe.

The starting schedule presents 16 training exposures under (.70,.30), followed by 32 probes. Half the participants retain the schedule; half change to (.30,.70) after probe 16. Alternative identity is independently counterbalanced. Use several motif families, counterbalanced conditions and exact sound outcomes where possible. Generate and save seeded realizations before responses; record intended probabilities and actual delivered exposure separately. Do not insert trials to force a desired individual learning effect. Freeze the state before eliciting the next response.

Let $\Delta q_i = \bar{q}_i(A)_{25:32} - \bar{q}_i(A)_{9:16}$. Estimate $\tau = E(\Delta q_i \mid \text{changed}) - E(\Delta q_i \mid \text{stable})$. A negative contrast supports exposure-dependent updating under this task. Fit participant/motif effects, baseline q , identity and realized exposure; retain assignment for intention-to-treat and report delivery-qualified sensitivity. Missingness and task-quality rules are independent of effect direction. For shortened schedules, corresponding windows and model-recovery tests must be frozen before confirmation.

Links to pleasure. In designated probes, obtain expected enjoyment V before the ending and pleasure Y afterwards, with shared 0–100 anchors. Counterbalance designated probes and response order across pilot variants to measure burden/reactivity. Report within-person/context associations between V , Y , delivered probability, realized exposure and acoustic change, with motif effects and appropriate bounded-outcome sensitivity. The controlled assignment tests exposure; a measured expectation–pleasure association is not itself a randomized causal effect of expectation. Sensory surprise, expected enjoyment and signed discrepancy $Y - V$ must not be collapsed[2, 3, 4].

Optional uncertainty assay. A binary distribution cannot freely vary entropy and target surprise as independent quantities. The former three-alternative design is retained only as a separate assay: scheduled contexts (.20,.75,.05) and (.20,.40,.40) hold target probability .20 while changing distributional entropy. Subjective uncertainty requires actual pre-outcome reports of the full probability vector, not the schedule or an MI score. If used, retain 24 training outcomes and 24 probes as starting counts, and preregister the assay, feasibility, power and multiplicity separately. A two-gesture probability allocation may be tested, but is not claimed qualified. Do not transfer this assay's entropy–surprise claim into the binary core[5].

The historical distribution floor .01, with .005/.02 sensitivities, applies only to the optional probability-vector calculations after normalization; report the floor and support. A scheduled target probability does not equal personal expectancy. If the optional assay is omitted, its earlier U-arm primary claim is omitted rather than relabelled as a pleasure correlation.

Additional assays. The value-learning rule $V_{t+1} = (1 - \alpha)V_t + \alpha Y_t$ can be compared with fixed value, repetition, unsigned surprise and reward-only alternatives using independent next-trial prediction. Correlating $Y_t - V_t$ with $V_{t+1} - V_t$ is algebraically coupled and is not validation. Resolved/delayed/withheld endings, separate wanting/liking blocks and regularity crossed with two-/four-item memory demands remain optional, separately consented designs[6, 7, 8]. Their absence limits SRP, DAED and WMED construct interpretation. Model outputs do not stand in for the missing assays.

S4. Mobile measurement and observation model

Gesture. During continuous transfer reporting, a relative vertical movement changes a bounded pleasure value. There is no page scrolling or momentum. A lifted finger leaves the selected value stationary; it does not automatically mean missing pleasure. Before scoring, require a deliberate initial placement and record the activation time. Invalid intervals include visibility loss, suspended audio, unqualified timing, lost state or incompatible delivery; a genuinely constant chosen value is valid. Practice, gain, anchors, sampling and device orientation rules are pilot-frozen. Keep a visually dark active page distinct from a locked phone.

Record value state at 5 Hz plus raw pointer events, audio start/stop, browser visibility and interruptions. Batch transport does not change local response timestamps. Collect an immediate overall rating after each excerpt, and separately recognition/unexpectedness. A retrospective overall rating may depend on peaks and ending; do not define it as the arithmetic mean of the trace. The current primary mean target is the mean of the valid aligned time samples; retrospective overall rating and trace peak are secondary.

Observation model. A useful specification is $y_{ise}(t) = g_i(\int_0^\infty K_i(u)\ell_{ise}(t-u)du) + \epsilon_{ise}(t)$, with latent experience-related process ℓ , reporting integration/delay K , scale use g and potentially autocorrelated error. One trace cannot jointly identify all these components without restrictions[9]. Estimate instrument delay with independent control tasks and assess transfer uncertainty, rather than claiming that a slider directly measures an internal neural signal. Freeze a common global or independently estimated participant-level lag rule in the pilot; apply it equally to all models, trim boundary samples and retain the raw trace.

Pilot lag candidates may span 0–3 seconds in 0.5-second steps, selected using independent instrument/calibration information and grouped validation. Report fixed nearby-lag sensitivities; never optimize each test trace's lag against the model being evaluated. Smoothing, if qualified, is fixed and causal for anticipation analyses. Temporal filters and all transitive feature support are recorded. Future-dependent features may be legal for a full-known-recording response predictor but illegal for a time-local listener-anticipation claim.

Pilot qualification. Use counterbalanced reporting/passive blocks with matched sound and post-excerpt overall ratings to estimate reporting reactivity. Separate motor/scale repeatability from repeat listening, which can itself change expectations and liking[10]. Record comprehension, completion, fatigue, missed intervals and device-specific timing. Starting qualification targets are at least 90% comprehension after permitted practice, at least 80% usable calibration/transfer completion among eligible consented pilot participants, no loss of acknowledged test events, and acceptable latency variability under the frozen analysis simulation. Report intervals and failures; these are proposed gates, not observed rates. Fix a burden target before examining MI gain. If core data fail qualification, redesign rather than claiming that nominally simple gestures guarantee low cognitive demand.

S5. Representation, personal learning and provenance

Let fixed audio representation $z_s(t) = F_\theta(x_s; t)$ feed a readout $\hat{y}_{ik} = g_\phi(z_{s_k}, w_{ik})$, with personal state $w_{ik} = \mathcal{U}_\phi(B_i, D_{i,<k})$. The engine θ and global learner ϕ are frozen before confirmation; personal state may change only under that learner using eligible earlier reports. Retain A: population, B: calibration-only and C: earlier-feedback-updated versions. B–A measures the history contribution; C–B measures updating. Their acoustic counterparts receive the same history and opportunities.

The starting mean readout uses two pilot-catalogue principal components per acoustic, temporal and selected-mechanism group; the baseline uses six acoustic components. A pilot-fitted population ridge model supplies shrinkage priors and $\lambda \in \{.1, 1, 10, 100\}$. Fit transformations and standardization within eligible training folds; freeze them before the outer evaluation. Continuous readouts use time-varying coordinates, the same allowed participant history and a separately qualified mapping, not a repeated constant mean prediction. Normalize each training excerpt's total weight so long traces do not dominate. Keep capacity-matched acoustic-plus-temporal and nonlinear alternatives as prespecified sensitivities with equal tuning budgets.

Personal fitting uses completed references and subsequently revealed trial outcomes; no parameters use a held-out target before its prediction. Save state before update, committed prediction, information cutoff, feature hashes and scoring mask. Mean and temporal models are fitted/scored at their corresponding scales. Background/expectation-informed additions use only actually measured variables; absent probability reports are missing, not inferred by naming an MI channel. The main model contrast uses C-MI versus C-acoustic. Other contrasts remain separate to avoid attributing extra history or readout flexibility to MI's representation.

Version inventory. The selected set contains HTP 12, ICEM 13, PWUP 10, UDP 10, SRP 19, PUPF 12, DAED 8, RPEM 8 and WMED 11 output coordinates, totalling 103. The 12 October static audit inspected 43 Python files and 58 citation entries including repeats. The source tree contained 89 module definitions with citation metadata; the older published architecture inventory had 84 mechanisms/889 coordinates. These are versioned inventories, not interchangeable runtime counts or validated constructs. The confirmatory manifest must pin the executed set and hashes rather than rely on either headline number.

Mechanism/source matrix. HTP operationalizes temporal hierarchy with weighted acoustic stability/shape features and an interaction; ICEM's information-content score uses acoustic variation rather than $-\log q(e)$ [11, 12, 2]. PWUP combines tonal/rhythmic proxies and applies a gate $E(1 - .5g)$; a source about uncertainty-dependent deviance responses does not establish this gate's empirical direction[13]. UDP combines acoustic and upstream uncertainty/confirmation branches, motivated by uncertain-music pleasure accounts[14]. PUPF combines acoustic proxies and SRP inputs; it is not the Cheung fitted chord-probability pleasure surface[5, 15].

SRP computes wanting/liking, tension and reward-related acoustic scores; DAED constructs anticipation/experience proxies from separate feature mixtures[16, 17, 18]. RPEM's centered current score is $\text{clip}(f_{03} - f_{04} + .5)$; it is not itself an observed $Y - V$ or an implemented participant-learning update[4]. WMED combines rhythm and memory-demand proxies derived from sound; it does not observe a participant's memory capacity or EEG entrainment[7, 8]. Preserve these computational definitions even when observations disagree. Citation existence, relevance, equation correspondence and empirical validation are separate checks.

Ablations and graphs. Record actual consumed dependencies, direct acoustic reads, fallbacks and temporal operators. Diagram edges do not prove runtime use; an unconsumed edge is a negative control. Remove a module's available outputs and refit under the same budget; test correlated groups and downstream paths explicitly. A code intervention establishes a computational change, not a human causal effect. Shuffled-person states, neutral names, matched acoustic changes and order/repetition controls distinguish plausible alternatives. A mechanism with a different observation target is not falsified by failing to explain a generic pleasure rating.

S6. Scoring, inference and independence

Eligibility and common support. Start with 24 transfer excerpts and require at least 18 valid excerpts for the primary participant-level model evaluation. A valid excerpt has at least 80% of the prespecified aligned window observed. The common mask is target/delivery-based and identical for competitors. Do not use forward filling to repair interrupted measurement or impute a primary target. A stationary valid instrument state is observed, not imputed. Retain all valid excerpts in a declared sensitivity analysis and report effects of the completion rule. Controlled-learning summaries initially require at least 75% valid probes and nonempty before/after windows; shortened schedules need a new frozen coverage rule.

Continuous score. For each eligible participant i and excerpt s with variable target $y_{is}(t)$, calculate Pearson $r_{is,m}$ using the same aligned samples for model m . Fisher-transform with clipping at $\pm(1 - 10^{-6})$, average across excerpts within person, then average across people and inverse-transform. This defines $r_{time,m}$. Each person has equal aggregate weight. A target with range below the pilot-qualified instrument resolution is not correlation-identifiable; retain it for mean/error outcomes and disclose its frequency. A variable target with a constant model prediction receives score zero by declared performance convention, counted as a model failure; its mathematical Pearson correlation remains undefined. Present a sensitivity using prediction-error scores without this convention. Eligibility never rewards a model for producing fewer variable predictions.

Mean score. Set $\bar{y}_{is}$ to the arithmetic mean over valid common-window samples and $\bar{\hat{y}}_{is,m}$ to the corresponding prediction mean. Calculate correlation across eligible excerpts within person, Fisher-average across people and inverse-transform to define $r_{mean,m}$. Participant-level nonvariation is handled analogously, with coverage and error metrics retained. This endpoint addresses ranking excerpt-mean experience within a person; a separate pooled or between-person correlation is descriptive only. Retrospective overall ratings are analyzed separately with their own denominators. Report all masks, score conventions, clipping and coverage.

Paired contribution. For each scale k , $\hat{\Delta}_k = r_{k,MI} - r_{k,acoustic}$. Resample participant and composition-family clusters with replacement, carrying all performances/exposures and recomputing both models' scores on the same bootstrap observations. The frozen predictions are scored, not returned. Use product cluster weights or equivalent pigeonhole resampling; sparse-family wild-bootstrap sensitivity and simulated type-I checks qualify the chosen implementation. Keep controlled motif identities separate from natural-music composition families.

For $B = 5000$ replicates, centered bootstrap errors are $e_{bk} = \hat{\Delta}_{bk} - \hat{\Delta}_k$. The proposed one-sided boundary test uses $p_k = [1 + \sum_b I(e_{bk} \geq \hat{\Delta}_k - .10)] / (B + 1)$. Verify calibration at the boundary under plausible crossed dependence; adopt a documented studentized alternative at pilot exit if required. Report bootstrap two-sided 95% intervals and joint/scaled effect estimates. The joint claim uses $p_M = \max(p_{time}, p_{mean})$, an intersection-union test requiring both gains. Apply Holm to p_M and the negative human updating contrast p_L . For human updating, a stratified randomization test respecting the actual assignment is preferred; fix its statistic, blocks and treatment of missingness before confirmation.

Point estimates above .10 with intervals crossing the boundary are promising but not confirmatory threshold evidence. Report no/weak/scale-specific/joint support explicitly. A positive gain smaller than .10 can still be informative but does not meet the chosen practical claim. The old one-point MAE threshold is preserved as a historical planning setting; MAE, MSE, calibration slope and signed bias are secondary diagnostics.

Secondary families. Separate personalization, policy value, mechanism and temporal-association families, declaring all tests and FDR procedures before confirmation. Fit participant/work mixed models for policy pleasure with order and prespecified familiarity controls. Policy scores reflect the released candidate pool. Avoid double-dipping in chosen peaks, time windows, module subsets or model classes. Model-selected pleasant moments are defined before human outcomes and compared with matched within/across-excerpt moments. Dense 5-Hz samples do not increase the independent participant count.

Generalization axes. New-response evaluation may share a person and composition; new-family evaluation excludes related compositions across fitting/test; new-person evaluation excludes that person's data from population training while permitting a specified personal calibration budget; new-performance evaluation holds passage family fixed. Crossed holdouts apply both person and family restrictions. Report these targets separately with shared outer folds across models. All feature selection, normalization, PCA, hyperparameters and lag choices respect training boundaries. Distinguish a first-exposure outcome from replay and a full-known-stimulus predictor from a past-only anticipation model.

S7. Feasibility, recovery and sample planning

The 60-registration pilot is an initial operational allocation rather than a guaranteed adequate validation sample. Exclude it from confirmatory inference. Evaluate candidate calibration doses and shorter task schedules through burden, coverage, comprehension and recovery. Preserve failed variants and the reason for selection. Pilot success is not measured by obtaining an attractive MI effect.

Simulate the full analysis with candidate analysable samples of 300/500/700/900, realistic family overlap, scale distributions, autocorrelation, participant effects, varying reliability and differential attrition. Nulls include no updating, acoustic-only prediction, nonlinear main effects without a distinct mechanism, correlated nuisance channels and gains at the .10 boundary. Positive alternatives span just below and meaningfully above .10 for both scales, plus human updating effects. Quantify Monte Carlo uncertainty, false-positive rates, interval coverage and power under the actual Holm rule. Target at least 90% power at a justified alternative strictly greater than the boundary; requiring 90% rejection at the boundary is invalid. Freeze the chosen alternative, final sample target and stopping rules without inspecting confirmatory data[19, 20].

The earlier standardized interaction .10 and probability-shift .10 were planning assumptions, not established effects. The optional three-choice assay needs its own recovery and power analysis. The registration cap allows at most 940 non-pilot registrations before attrition; it does not ensure 940 analysable participants. If available precision is inadequate, amend recruitment/scope or state the limitation rather than redefine success after seeing outcomes. Predeclare infrastructure-pausing and attrition accounting; withdrawals and failed baselines do not silently reopen cap slots.

S8. Acquisition, records and operational acceptance

Research objects and permissions. A versioned release binds protocol, consent, catalogue, audio preprocessing, engine, selected coordinates, personal calibration, readout coefficients, random seeds, trial schedule and hidden predictions. Store quantities with namespace, units, temporal support and status: acoustic feature, model score, probability report, pleasure report or later prediction. Keep raw observations intact. The browser receives only its authorized participant-facing manifest and short-lived audio access; hidden predictions, other profiles and server credentials stay outside it. Separate contact records, consent records and pseudonymous research views.

Preparation. A shared worker decodes/validates each licensed audio asset and computes features once per digest. A personal worker consumes a locked calibration, initializes models, assigns conditions/policies and serializes the release. A proposed durable Postgres-backed queue may implement bounded leases and at-least-once delivery[21]; no provider deployment is claimed. Baseline locking and queue/outbox creation share a transaction. A compare-and-swap version check rejects stale jobs after corrected calibration, withdrawn consent or removed rights. Completed obsolete computation may be retained for audit but cannot authorize listening.

Registration states distinguish verified, consented, calibration-in-progress/locked, preparation-queued/running, ready, active and completed, with ineligible, failed, expired and withdrawn branches. Only server-authorized transitions count. Readiness validates the pilot-frozen route/coverage rules, quotas, permitted assets and exact model/configuration version. Duplicate movement/remaster references are checked by musical identity, not only byte hashes. The current three-source rules replace former requirements that every reference be previously known.

Delivery and clocks. Pin audited experiment components around Web Audio, for example jsPsych[22], with explicit mobile qualification. Preload/decode only approved assets. Record audio-context start, browser-monotonic event time, server receipt, visibility, underruns and device changes. Estimate audio/browser mapping and drift on representative hardware; physical validation is separate from network latency[23]. Timezone/wall-clock changes never reorder exposure events. A device/browser combination lacking required guarantees is ineligible until qualified.

Durability. Every event has an exposure-local sequence, stable identifier, schema version and payload hash. Ten 5-Hz value samples form a two-second batch, preserving sample timestamps. Acknowledgement follows database commit. Identical identifier/payload retries return the original receipt; conflicts are rejected and logged. A bounded local queue survives temporary network loss within storage/authorization constraints. Expired credentials, exhausted storage or missing audio stop at a safe boundary. Reload reconciles receipts and pending events without guessing whether unrecorded sound was heard. A replay or partial restart creates a linked new exposure, never an overwrite.

Synthetic acceptance. Exercise the full registration-to-export path before a human pilot. Race 1,000-slot claims with retries: at most 1,000 distinct accepted registrations. Inject duplicate jobs, lease expiry, calibration changes, withdrawal and rights removal: no stale release may authorize playback. Drop acknowledgements, reorder batches and send conflicting payloads: fixture events must be reconstructed exactly with explicit errors, no silent duplicates or loss of acknowledged data. Test dark-active screens, lock/suspension, audio interruptions and storage exhaustion on supported phones.

Load-test 50 sessions for 30 minutes and 100 for 15 minutes as starting engineering fixtures, with representative payloads and

authorization. Proposed gates remain p95 commit acknowledgement below one second, zero lost acknowledged events and less than .1% errors excluding intentional invalid requests. Measure audio/response timing independently. Network inspection must show no live MI/GPU/LLM inference during prepared sessions. Restore a synthetic snapshot including later withdrawal, reconcile permissions and reproduce earlier-only replay scores. These are proposed acceptance tests, not reported passes or service guarantees.

S9. Audit, reproducibility and analysis assistance

An analysis bundle contains immutable observation release, exclusions, variable definitions, dependency/support records, source roles, executable program, environment versions, randomness, masks, fold assignments, estimates and diagnostics. A conclusion maps to the exact numerical artifact and qualifying observations. Hashes establish identity, not psychological validity or anonymity. Keep expected and observed quantities separate, and record source existence, relevance and support independently.

An LLM may assist specification, code drafting or literature organization. The evaluated learner and analysis are saved programs; a newly generated answer is not a reproduction of yesterday's calculation. Validate outputs against known-truth fixtures, independent numerical checks and information-access tests. Freeze analyst instructions, tools and access when comparing analyst performance. A fluent explanation or successful execution is insufficient evidence of an adequate scientific conclusion.

Preserve admissible alternative analyses with reasons and identifiers; do not select a favourable one after seeing targets[24, 25]. Use a claim ledger with target, estimate, uncertainty, observation level, supporting artifact and limitation. Registered and exploratory analyses are labelled separately[20]. The main study's saved computations remain reproducible without a live LLM service. The LLM benchmark in S11 measures a different claim and has separate data and evaluation budgets.

S10. Interpretation, revision, access and disclosure

Report the human and model decisions independently. Weak MI gain can motivate inspection of measurement, input mappings, temporal support, shared features and readout before proposing a mechanism change. A source may correctly motivate a relation while an acoustic implementation fails to represent it. A sensitive, faithful test that challenges a source-derived prediction can refine the literature; a failed operational proxy alone cannot. A predictive model may be useful while compatible with several mechanisms[26].

After the primary freeze, mechanism updates define MI v2 with a new manifest and change rationale. Use fresh people/materials or a genuinely untouched reserved evaluation set. Discovery data used to select a revision are development data for that revision. Preserve the original v1 predictions and failures. Biological interpretations require observation-specific tests: behaviour, EEG, MEG and fMRI are not interchangeable outcomes.

Recruitment requires institutional responsibility, ethics review, plain-language consent, compensation, recording permissions and withdrawal procedures. State retention/access rules and limitations after irreversible aggregation. Minimize contact/demographic collection and distinguish account verification from consent. Release approved code, synthetic fixtures, permitted catalogue metadata and sufficiently de-identified summaries, with controlled access to restricted material[27]. A public protocol is not a launched experiment; an email-interest form is not enrollment or consent. Funding, affiliation and author contributions must be accurate at submission. AI assistance is disclosed as editorial/programming support rather than an empirical analyst validation.

S11. Separate future studies retained from the earlier framework

Documentation and LLM benchmark. The independent question is whether operational definitions and verified provenance improve an analyst's scientific conclusions when numerical arrays are fixed. D0 supplies prespecified acoustic features; D1 supplies MI values and a minimal dictionary; D2 adds operational definitions and temporal support; D3 adds executed dependencies, fallbacks and verified source roles. Common covariates, allowable outcomes, timing and resources stay identical. D3–D1 is the primary documentation contrast with MI arrays and analyst fixed. D2–D1 and D3–D2 diagnose definitions and provenance. D1–D0 tests a representation package, not each mechanism, and requires matched-dimension/tuning sensitivities.

Retain the proposed inventory of 24 development tasks and 120 evaluation tasks, 30 each for known-truth recovery, null/confound diagnosis, accessible-analysis replication and held-out prediction/discovery. Four conditions with three fresh-context repetitions produce 1,440 main runs before controls. These are planning counts, not a power justification or executed benchmark. Use newly generated instances/prospective observations where training contamination matters; keep reference labels/seeds outside analyst permissions. Existing precedents motivate tool-use evaluation without validating MI-Lab[28, 29, 30, 31, 32, 33].

Score mean success rather than best attempt. A task passes only when its declared numerical target, information boundary, rerunnable artifact and supported conclusion all pass. Appropriate negative or qualified answers can pass underidentified/null tasks; universal refusal cannot. Include timeouts and failed repairs with equal opportunities and costs. Two blinded reviewers adjudicate support and retain disagreements. LLM-only grading is insufficient. Include verbosity-matched documentation, neutral/mismatched labels, unused-edge controls and an expert-authored frozen pipeline. Changing provider/model is a separate analyst comparison.

For task q , condition c and repetition r , let s_{qcr} be binary success and $\bar{s}_{qc}$ the repetition mean. Estimate the equal-task-weighted difference between D3 and D1, stratified by task family. Resample independent task/source families, preserving paired conditions; repetitions are not independent new tasks. Plan precision/power with dependence and failure rates before evaluation. Secondary outcomes include numerical error, rerun agreement, unauthorized access, unsupported biological claims, citation support, expert correction time, latency and total computational cost. Do not conceal an increase in unsupported claims inside an aggregate success score.

The benchmark has its own protocol, evaluation freeze and publication decision. It is deferred from the human-study primary claims; its full earlier framework remains archived. The human study may supply later eligible tasks, but reused observations and analyst exposure must be declared. No benchmark performance is asserted here.

Independent neural extension. A later EEG/MEG/fMRI study requires independent acquisition, physical audio synchronization, qualified markers and device-delay/drift handling. Retain recording/exposure identity while specifying modality-specific resolution, preprocessing and nuisance models. An illustrative observation family is $b_{iv}(t) = [\mathcal{H}_v * f_{iv}(z_s, c_{ise})](t) + \eta_{iv}(t)$, with measured activity b , fixed features z , admissible context c and modality operation $\mathcal{H}$. fMRI haemodynamics and EEG/MEG temporal/forward models differ from the slider kernel. BIDS supports identity and metadata, not biological validation[34].

Require held-out neural prediction against acoustic/timing/motor/nuisance alternatives, same-capacity readouts and predeclared regions/windows. Estimate reliability/noise ceilings under explicit assumptions. Correlated features may prevent unique attribution. Work on non-robust LLM–brain alignment motivates controls for shuffled splits, autocorrelation, position and envelope; it is not an MI neural result[35]. Controlled stimulus manipulations should make alternative explanations disagree. No neural region, transmitter or working-memory claim is established by the browser data alone.

Appendix A. Historical figure record

Archival plates from 11 October 2026. Settings such as 20 familiar works, desktop eligibility, three-choice U/L assignment and the previous primary endpoint are superseded where S1 states. These are not the current recruitment or analysis contract.

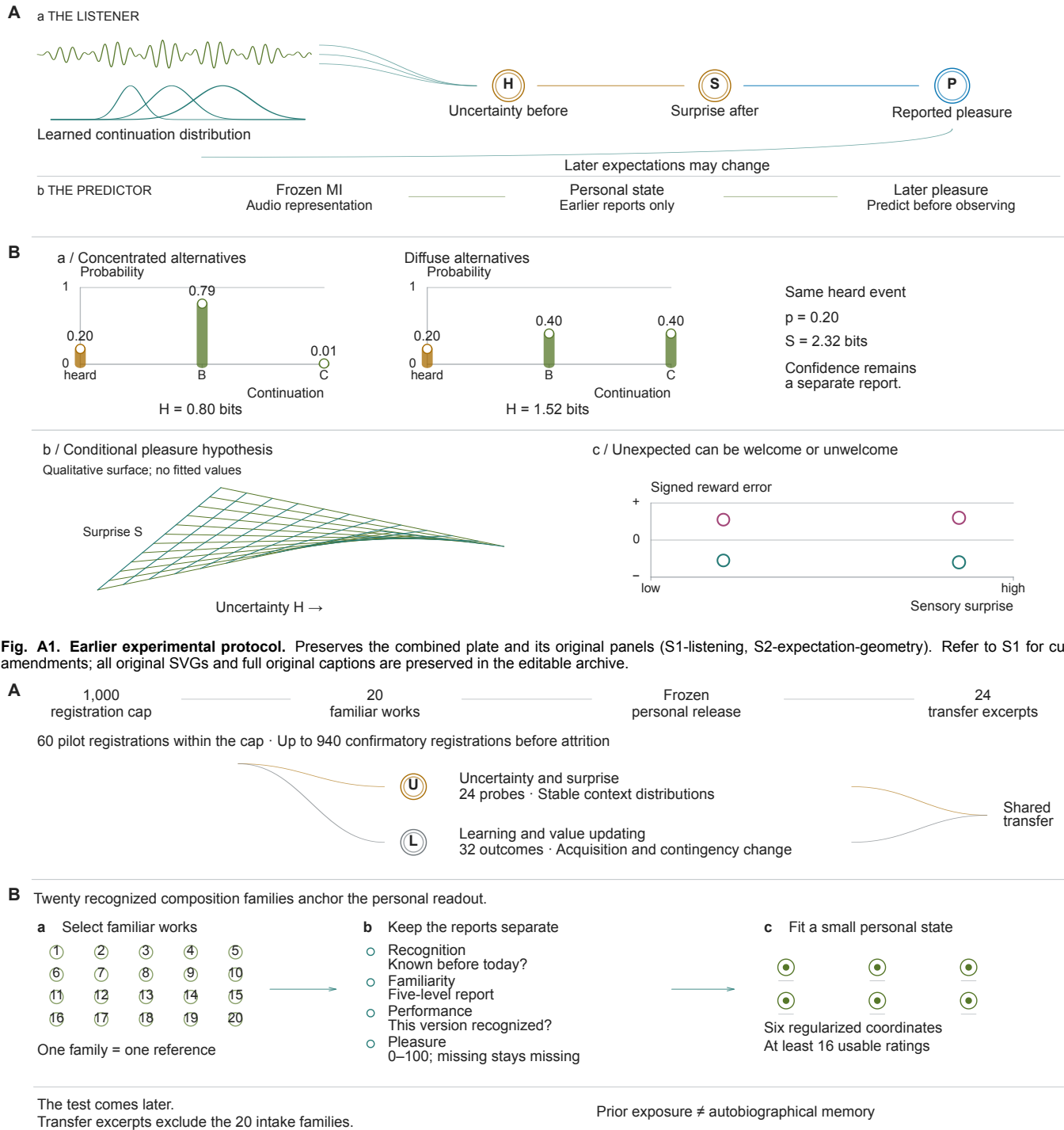


Fig. A2. Earlier experimental protocol. Preserves the combined plate and its original panels (S3-design, S8-personal-intake). Refer to S1 for current amendments; all original SVGs and full original captions are preserved in the editable archive.

Appendix A. Historical figure record

Archival plates from 11 October 2026. Settings such as 20 familiar works, desktop eligibility, three-choice U/L assignment and the previous primary endpoint are superseded where S1 states. These are not the current recruitment or analysis contract.

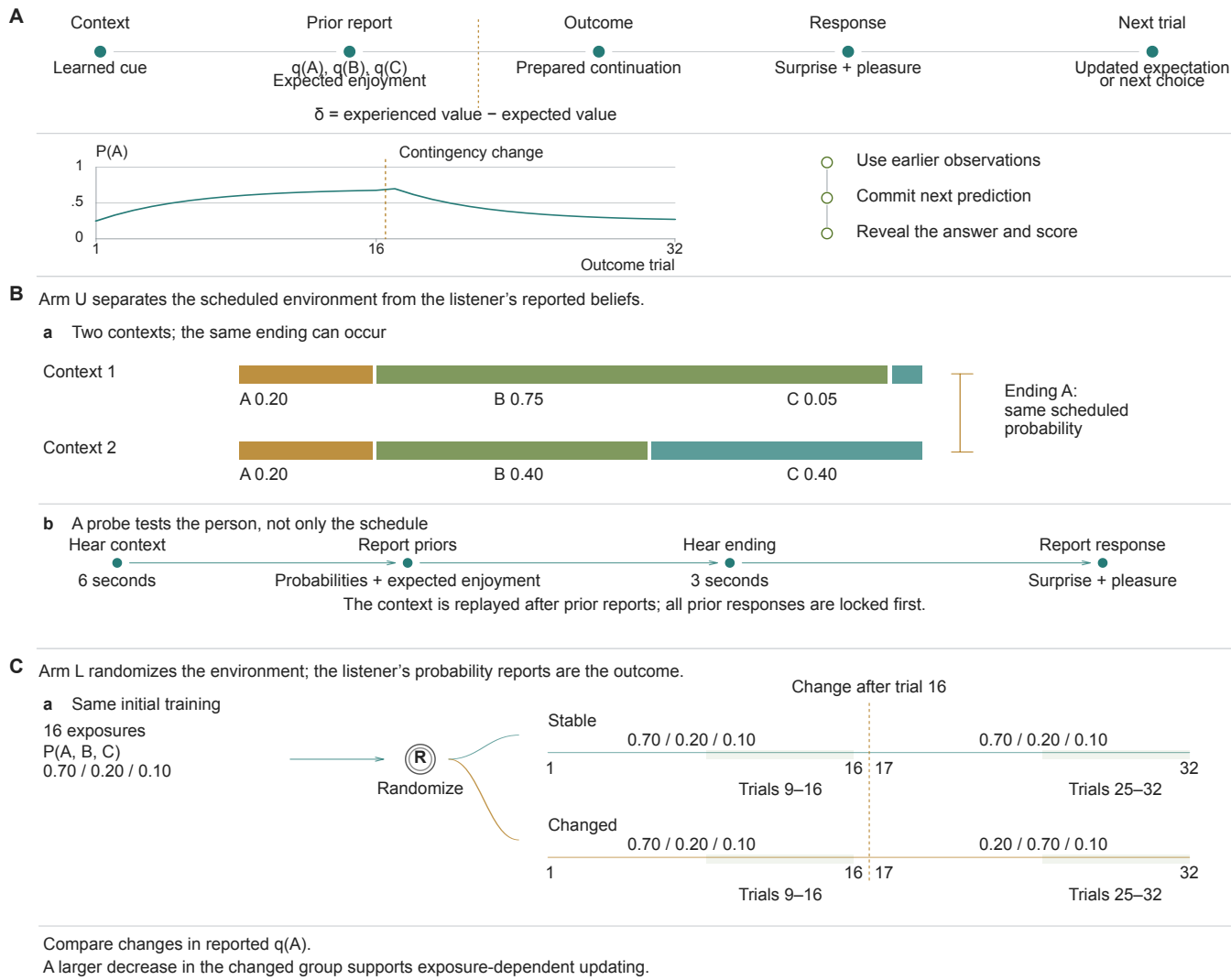


Fig. A3. Earlier experimental protocol. Preserves the combined plate and its original panels (S4-learning, S9-context-experiment, S10-learning-experiment). Refer to S1 for current amendments; all original SVGs and full original captions are preserved in the editable archive.

Appendix A. Historical figure record

Archival plates from 11 October 2026. Settings such as 20 familiar works, desktop eligibility, three-choice U/L assignment and the previous primary endpoint are superseded where S1 states. These are not the current recruitment or analysis contract.

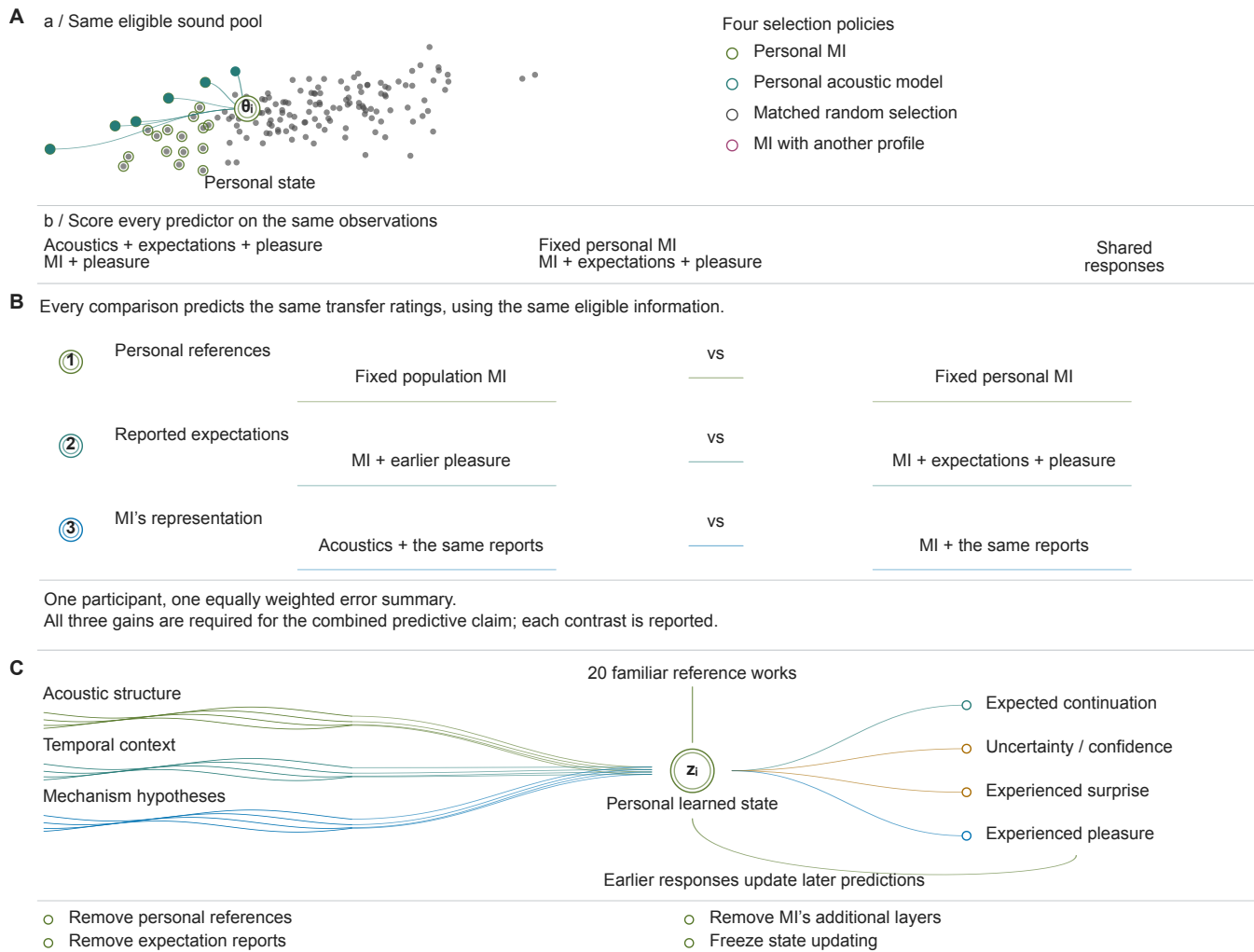


Fig. A4. Earlier experimental protocol. Preserves the combined plate and its original panels (S5-transfer, S11-predictive-increments, S7-integrated). Refer to S1 for current amendments; all original SVGs and full original captions are preserved in the editable archive.

Appendix A. Historical figure record

Archival plates from 11 October 2026. Settings such as 20 familiar works, desktop eligibility, three-choice U/L assignment and the previous primary endpoint are superseded where S1 states. These are not the current recruitment or analysis contract.

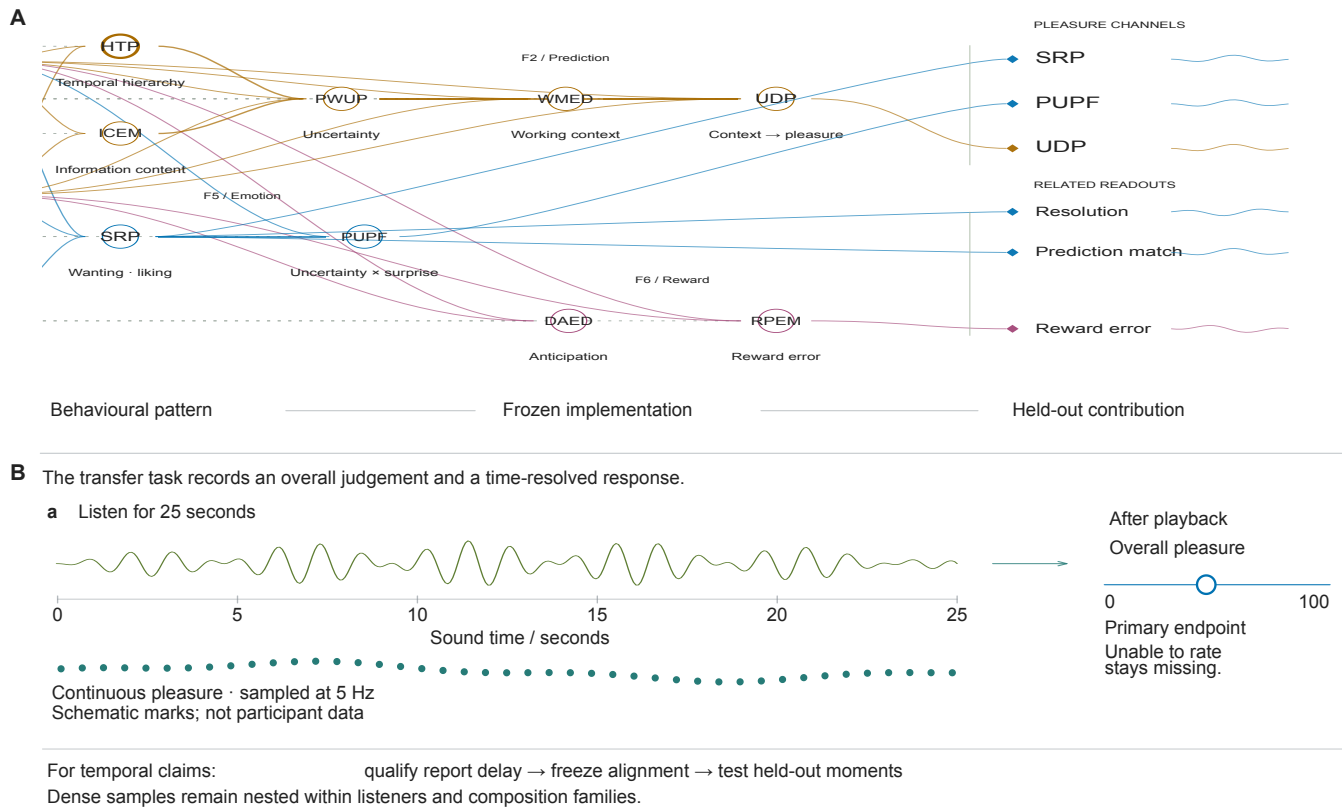


Fig. A5. Earlier experimental protocol. Preserves the combined plate and its original panels (S6-mechanisms, S12-response-measurement). Refer to S1 for current amendments; all original SVGs and full original captions are preserved in the editable archive.

A Interpret the behavioural relationship and the corresponding MI account separately.

	MI agreement / predictive contribution	
	Supported	Not supported precisely
Behavioural pattern supported	Converging evidence Still test alternative explanations.	Locate a model limitation The human relationship can stand.
Behavioural pattern not supported precisely	Prediction without the account Useful prediction; mechanism unresolved.	Revise the account and task Report population and task limits.

A wide interval is inconclusive; it is not evidence of absence.

Fig. A6. Earlier experimental protocol. Preserves the combined plate and its original panels (S13-interpretation). Refer to S1 for current amendments; all original SVGs and full original captions are preserved in the editable archive.

Appendix A. Historical figure record

Archival plates from 11 October 2026. Settings such as 20 familiar works, desktop eligibility, three-choice U/L assignment and the previous primary endpoint are superseded where S1 states. These are not the current recruitment or analysis contract.

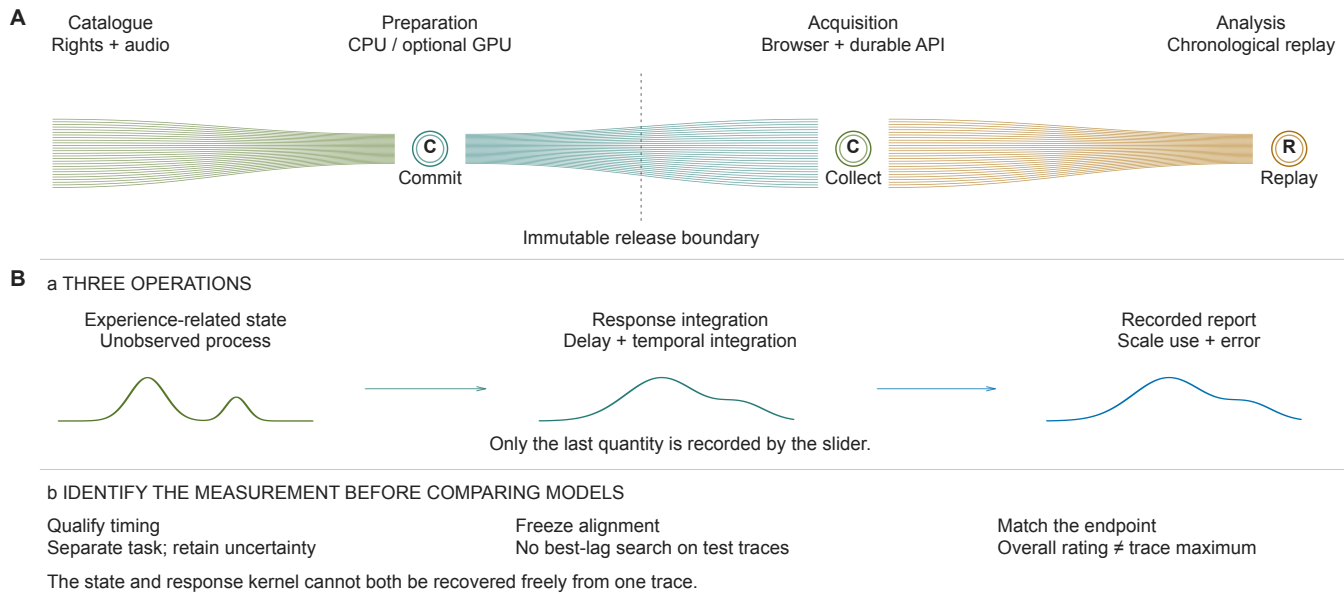


Fig. A7. Earlier measurement framework. Preserves the combined plate and its original panels (T1-system, N1-observation-model). Refer to S1 for current amendments; all original SVGs and full original captions are preserved in the editable archive.

Appendix A. Historical figure record

Archival plates from 11 October 2026. Settings such as 20 familiar works, desktop eligibility, three-choice U/L assignment and the previous primary endpoint are superseded where S1 states. These are not the current recruitment or analysis contract.

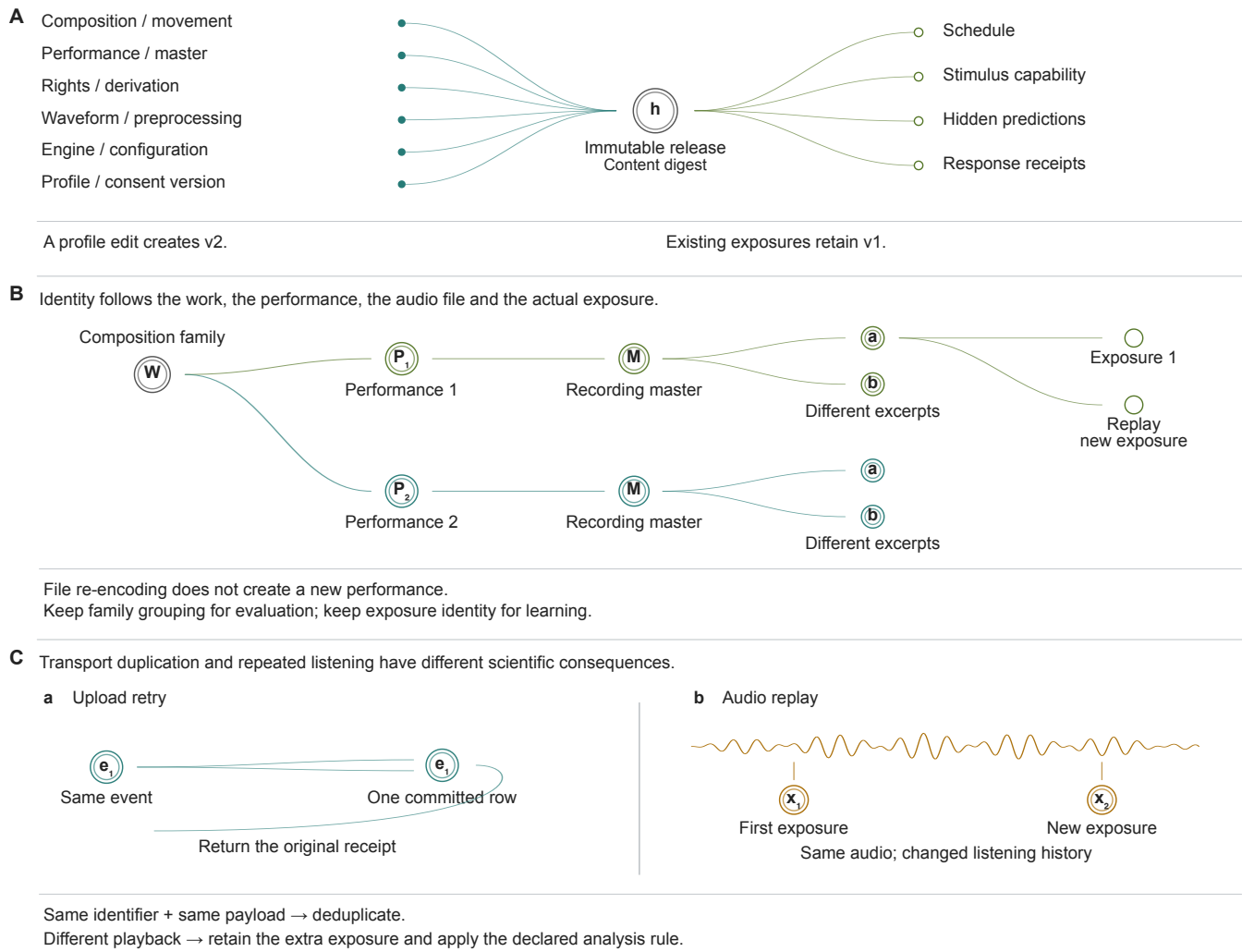


Fig. A8. Earlier measurement framework. Preserves the combined plate and its original panels (T2-provenance, T6-musical-identity, T8-retry-exposure). Refer to S1 for current amendments; all original SVGs and full original captions are preserved in the editable archive.

Appendix A. Historical figure record

Archival plates from 11 October 2026. Settings such as 20 familiar works, desktop eligibility, three-choice U/L assignment and the previous primary endpoint are superseded where S1 states. These are not the current recruitment or analysis contract.

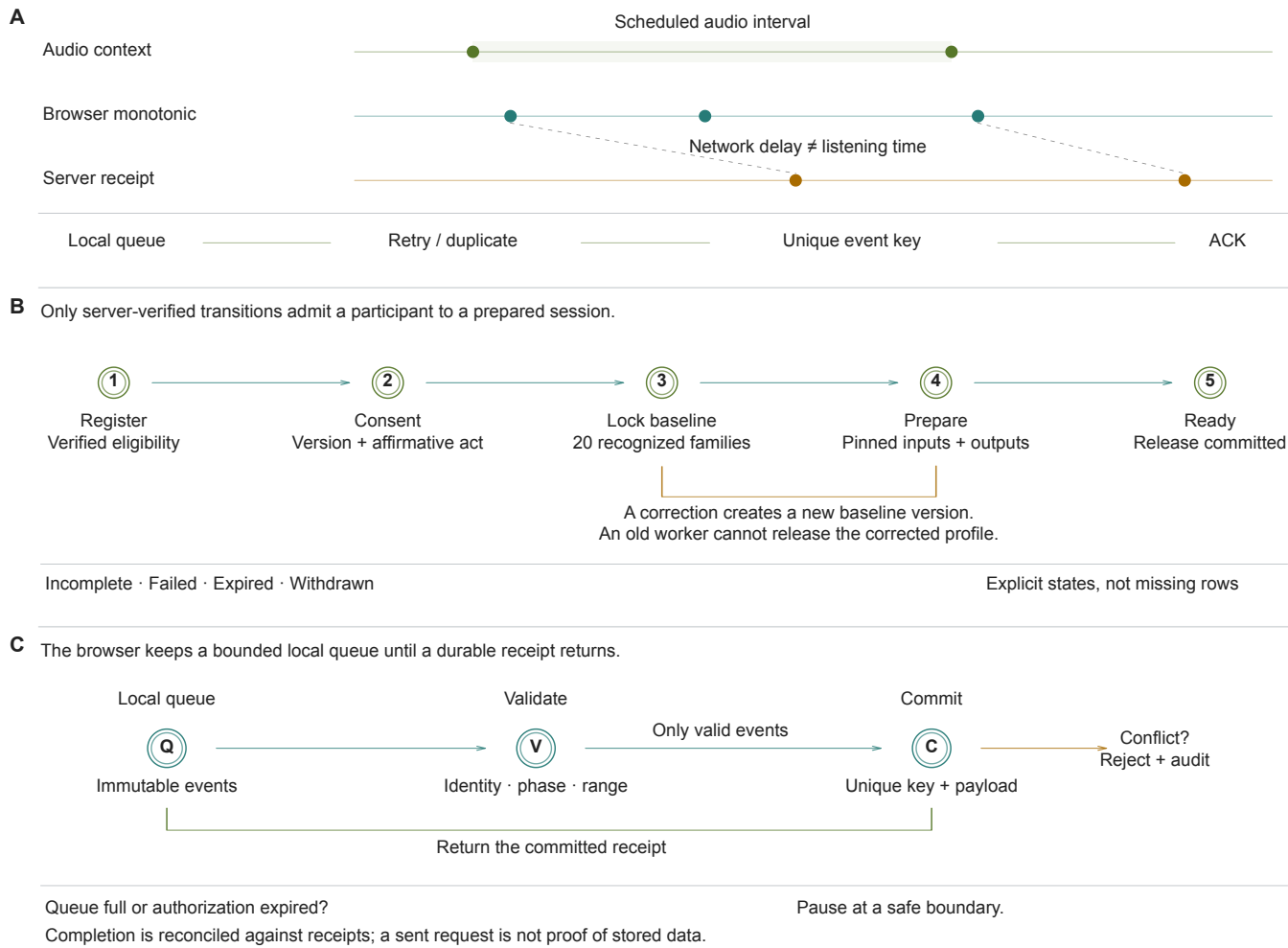
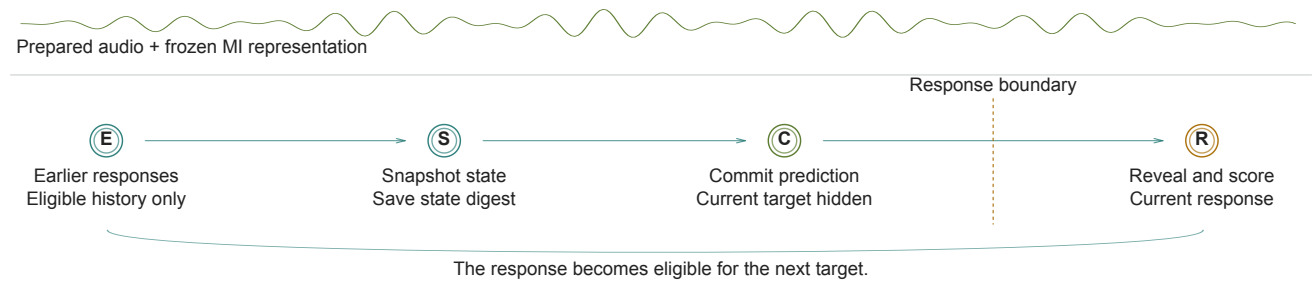


Fig. A9. Earlier measurement framework. Preserves the combined plate and its original panels (T3-acquisition, T7-readiness, T9-commit-receipt). Refer to S1 for current amendments; all original SVGs and full original captions are preserved in the editable archive.

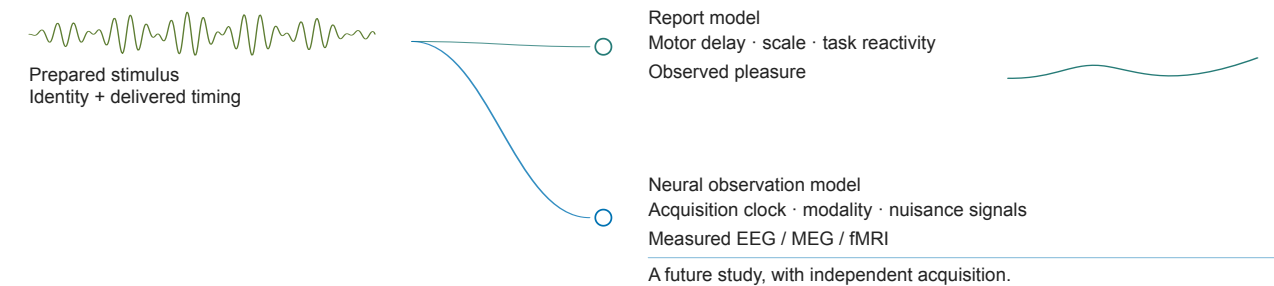
Appendix A. Historical figure record

Archival plates from 11 October 2026. Settings such as 20 familiar works, desktop eligibility, three-choice U/L assignment and the previous primary endpoint are superseded where S1 states. These are not the current recruitment or analysis contract.

A Chronological replay tests a frozen learner without exposing its next answer.



B Same versioned sound · known exposure history · independently measured outcomes



Held-out neural prediction must beat matched acoustic and temporal alternatives.
A brain correlation does not identify a unique biological mechanism.

Fig. A10. Earlier measurement framework. Preserves the combined plate and its original panels (T10-chronological-replay, N2-neural-observation). Refer to S1 for current amendments; all original SVGs and full original captions are preserved in the editable archive.

Appendix A. Historical figure record

Archival plates from 11 October 2026. Settings such as 20 familiar works, desktop eligibility, three-choice U/L assignment and the previous primary endpoint are superseded where S1 states. These are not the current recruitment or analysis contract.

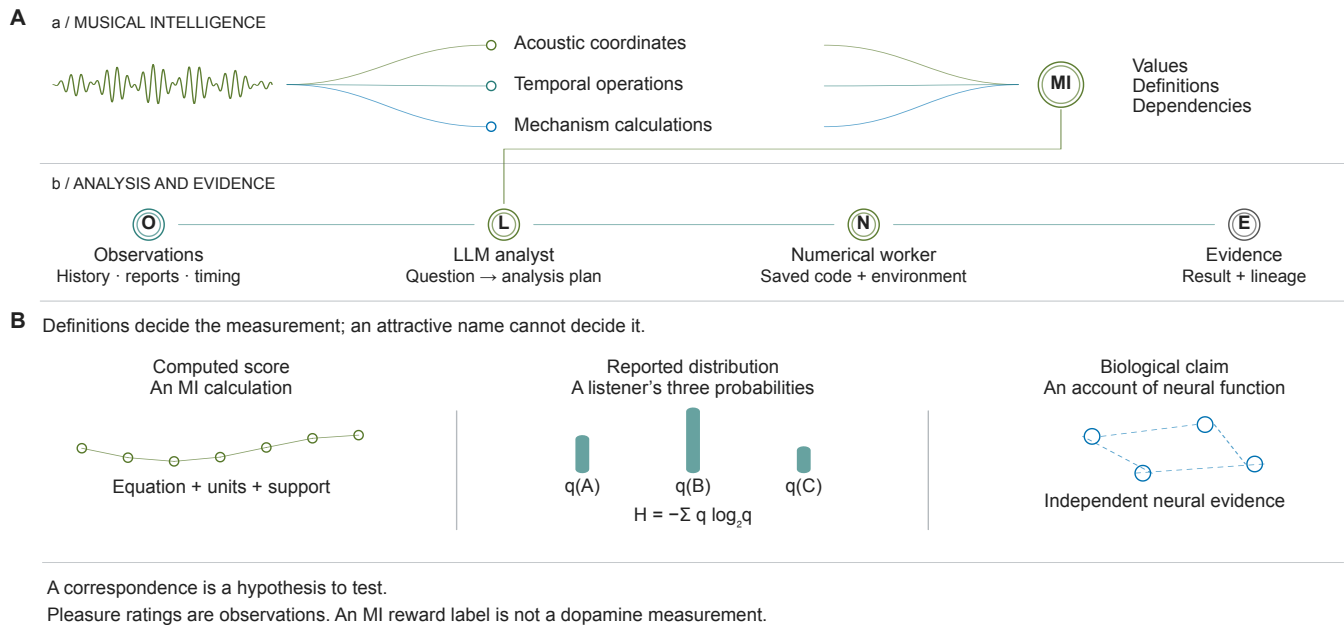


Fig. A11. Earlier inference framework. Preserves the combined plate and its original panels (F1-scientific-lens, F6-evidence-types). Refer to S1 for current amendments; all original SVGs and full original captions are preserved in the editable archive.

Appendix A. Historical figure record

Archival plates from 11 October 2026. Settings such as 20 familiar works, desktop eligibility, three-choice U/L assignment and the previous primary endpoint are superseded where S1 states. These are not the current recruitment or analysis contract.

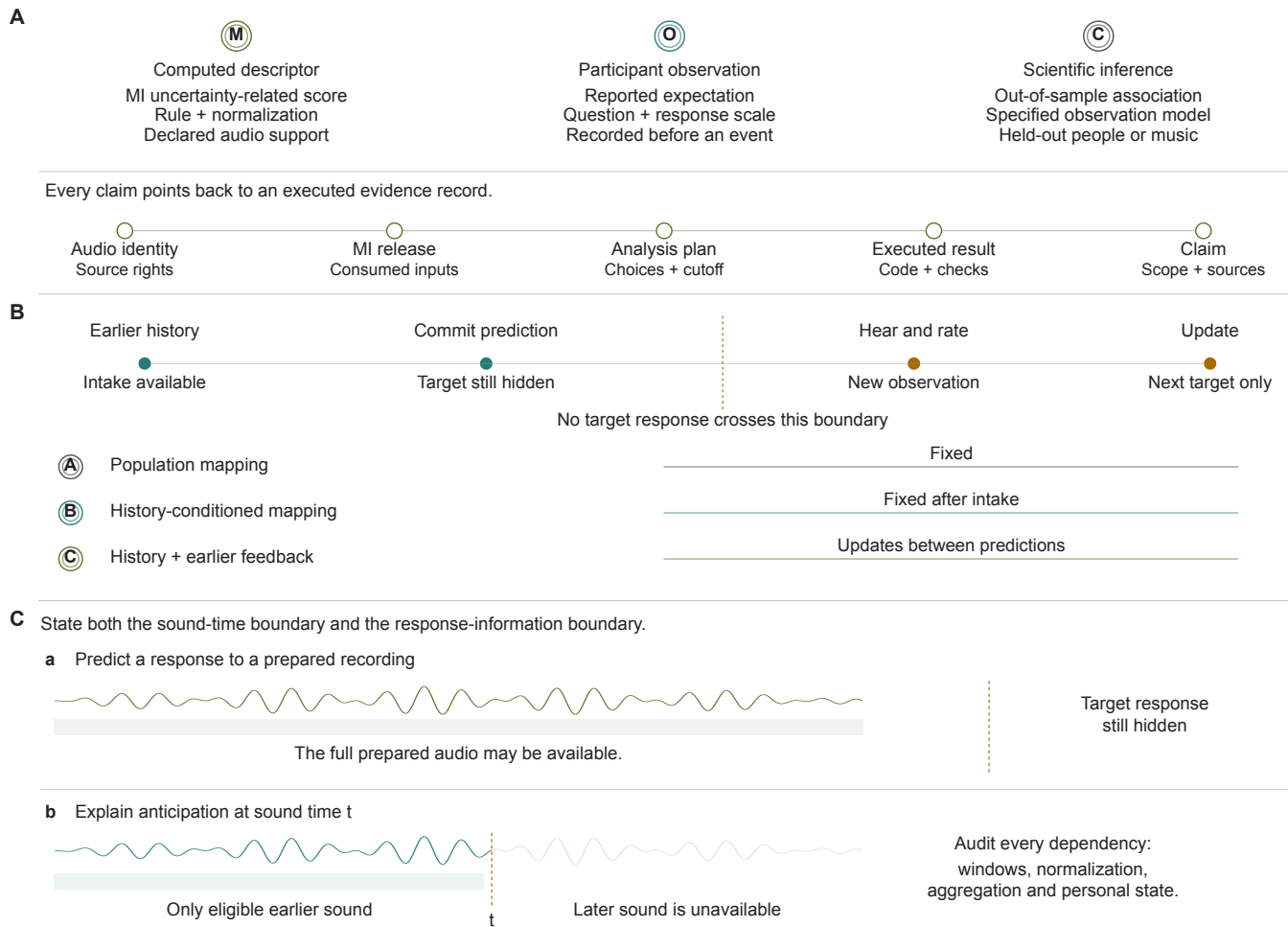


Fig. A12. Earlier inference framework. Preserves the combined plate and its original panels (F2-evidence-contract, F3-information-boundary, F7-two-boundaries). Refer to S1 for current amendments; all original SVGs and full original captions are preserved in the editable archive.

Appendix A. Historical figure record

Archival plates from 11 October 2026. Settings such as 20 familiar works, desktop eligibility, three-choice U/L assignment and the previous primary endpoint are superseded where S1 states. These are not the current recruitment or analysis contract.

A The LLM helps specify and write the analysis; a retained program produces the numbers.



B Choose the validation split to match the claim, and use the same folds for competitors.

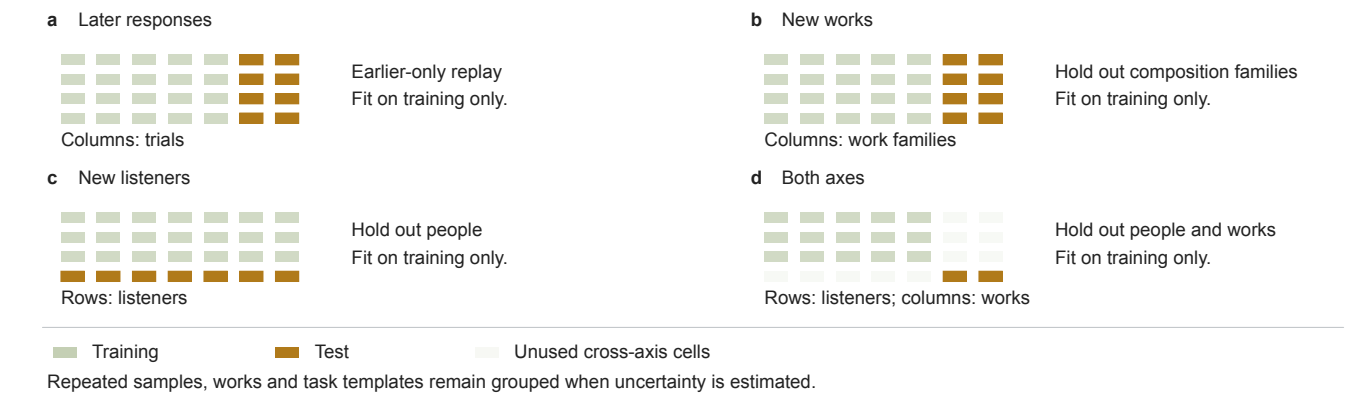


Fig. A13. Earlier inference framework. Preserves the combined plate and its original panels (F8-executed-evidence, F9-generalization). Refer to S1 for current amendments; all original SVGs and full original captions are preserved in the editable archive.

Appendix A. Historical figure record

Archival plates from 11 October 2026. Settings such as 20 familiar works, desktop eligibility, three-choice U/L assignment and the previous primary endpoint are superseded where S1 states. These are not the current recruitment or analysis contract.

A A better hybrid result does not reveal which component helped.

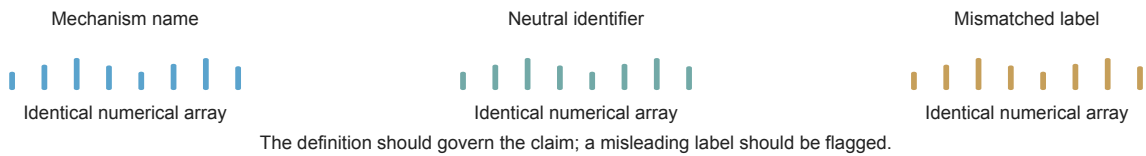
①	Representation	Acoustic features ↔ MI features Hold observations, learner and fitting resources fixed.
②	Documentation	MI values ↔ MI values + evidence contract Hold numerical arrays, analyst and permissions fixed.
③	Analyst	One model or budget ↔ Another model or budget Hold task, representation and evidence access fixed.

D3 – D1 isolates the evidence contract; it does not establish MI's acoustic superiority.

D0	D1	D2	D3
Acoustic	MI values	MI + contract	MI + contract + lineage
● Numeric representation ○ Expanded definitions ○ Dependency / source	● Numeric representation ○ Expanded definitions ○ Dependency / source Same analyst · Same tools and budget · Same admissible observations	● Numeric representation ● Expanded definitions ○ Dependency / source	● Numeric representation ● Expanded definitions ● Dependency / source
Correct Numerical acceptance tests	Reproducible Independent execution	Calibrated Null cases + uncertainty	

C Prespecified benchmark controls expose persuasive names and unconsumed graph edges.

a Same values and definitions, different labels



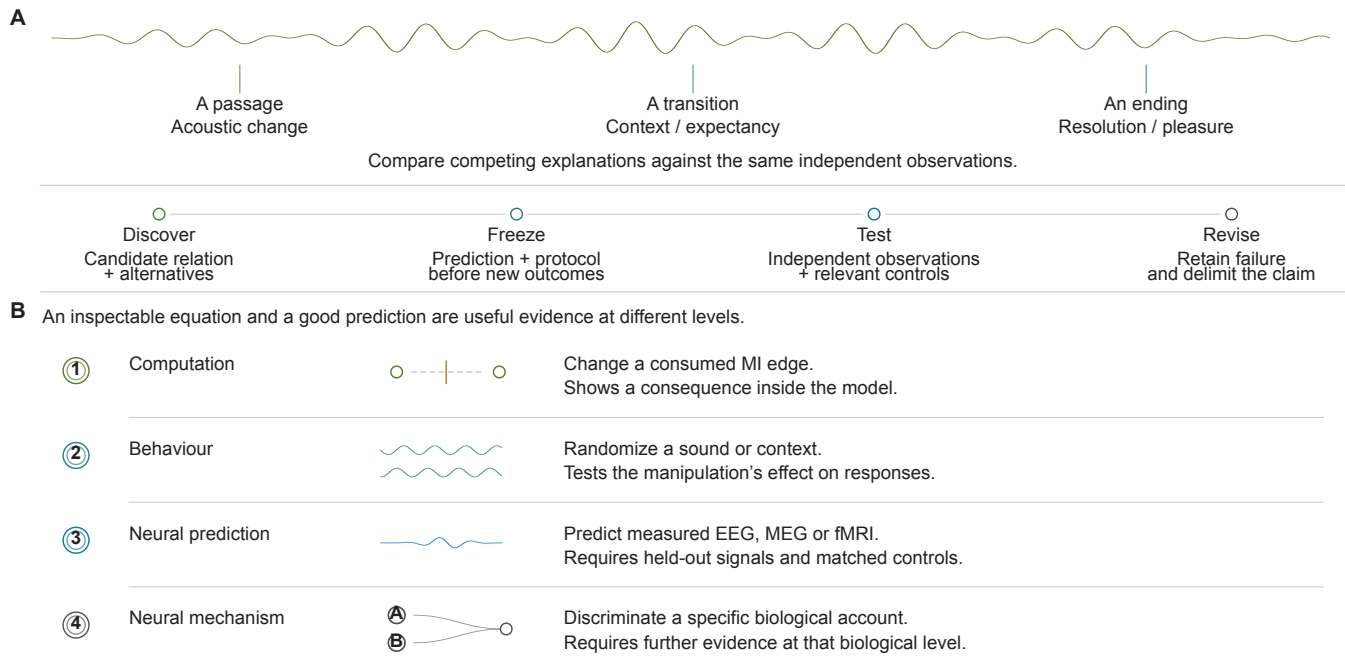
b A drawn edge matters only if the computation used it



Fig. A14. Earlier inference framework. Preserves the combined plate and its original panels (C1-attribution, F4-benchmark, F10-label-controls). Refer to S1 for current amendments; all original SVGs and full original captions are preserved in the editable archive.

Appendix A. Historical figure record

Archival plates from 11 October 2026. Settings such as 20 familiar works, desktop eligibility, three-choice U/L assignment and the previous primary endpoint are superseded where S1 states. These are not the current recruitment or analysis contract.



No automatic promotion from a channel name to a brain mechanism.

Fig. A15. Earlier inference framework. Preserves the combined plate and its original panels (F5-discovery-loop, F11-brain-claims). Refer to S1 for current amendments; all original SVGs and full original captions are preserved in the editable archive.

References

- [1] Hawthorne, C. et al. Enabling factorized piano music modeling and generation with the MAESTRO dataset. *ICLR* (2019). Dataset, versions and licence.
- [2] Egermann, H., Pearce, M. T., Wiggins, G. A. & McAdams, S. Probabilistic models of expectation violation predict psychophysiological emotional responses to live concert music. *Cognitive, Affective, & Behavioral Neuroscience* **13**, 533–553 (2013). doi:10.3758/s13415-013-0161-y.
- [3] Gold, B. P. et al. Musical reward prediction errors engage the nucleus accumbens and motivate learning. *Proceedings of the National Academy of Sciences USA* **116**, 3310–3315 (2019). doi:10.1073/pnas.1809855116.
- [4] Gold, B. P. et al. Auditory and reward structures reflect the pleasure of musical expectancies during naturalistic listening. *Frontiers in Neuroscience* **17**, 1209398 (2023). doi:10.3389/fnins.2023.1209398.
- [5] Cheung, V. K. M. et al. Uncertainty and surprise jointly predict musical pleasure and amygdala, hippocampus, and auditory cortex activity. *Current Biology* **29**, 4084–4092.e4 (2019). doi:10.1016/j.cub.2019.09.067.
- [6] Lehne, M., Rohrmeier, M. & Koelsch, S. Tension-related activity in the orbitofrontal cortex and amygdala: an fMRI study with music. *Social Cognitive and Affective Neuroscience* **9**, 1515–1523 (2014; online 2013). doi:10.1093/scan/nst141.
- [7] Noboa, M. de L., Kertész, C. & Honbolygó, F. Neural entrainment to the beat and working memory predict sensorimotor synchronization skills. *Scientific Reports* **15**, 10466 (2025). doi:10.1038/s41598-025-93948-9.
- [8] Yuan, Y. et al. Decoding auditory working memory load from EEG alpha oscillations. *Psychophysiology* **62**, e70210 (2025). doi:10.1111/psyp.70210.
- [9] Flake, J. K. & Fried, E. I. Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science* **3**, 456–465 (2020). doi:10.1177/2515245920952393.
- [10] Albury, A. W., Bianco, R., Johnson, A. P. & Penhune, V. B. Stimulus predictability and liking enhance auditory–motor encoding and memory for melodies. *Annals of the New York Academy of Sciences* **1560**, e70310 (2026). doi:10.1111/nyas.70310.
- [11] Norman-Haignere, S. V. et al. Multiscale temporal integration organizes hierarchical computation in human auditory cortex. *Nature Human Behaviour* **6**, 455–469 (2022). doi:10.1038/s41562-021-01261-y.
- [12] Bonetti, L. et al. Spatiotemporal brain hierarchies of auditory memory recognition and predictive coding. *Nature Communications* **15**, 4313 (2024). doi:10.1038/s41467-024-48302-4.
- [13] Quiroga-Martinez, D. R. et al. Reduced prediction error responses in high- as compared to low-uncertainty musical contexts. *Cortex* **120**, 181–200 (2019). doi:10.1016/j.cortex.2019.06.010.
- [14] Mencke, I., Omigie, D., Wald-Fuhrmann, M. & Brattico, E. Atonal music: can uncertainty lead to pleasure? *Frontiers in Neuroscience* **12**, 979 (2019). doi:10.3389/fnins.2018.00979.
- [15] Gold, B. P. et al. Predictability and uncertainty in the pleasure of music: a reward for learning? *Journal of Neuroscience* **39**, 9397–9409 (2019). doi:10.1523/JNEUROSCI.0428-19.2019.
- [16] Salimpoor, V. N. et al. Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. *Nature Neuroscience* **14**, 257–262 (2011). doi:10.1038/nn.2726.
- [17] Salimpoor, V. N. et al. Interactions between the nucleus accumbens and auditory cortices predict music reward value. *Science* **340**, 216–219 (2013). doi:10.1126/science.1231059.
- [18] Ferreri, L. et al. Dopamine modulates the reward experiences elicited by music. *Proceedings of the National Academy of Sciences USA* **116**, 3793–3798 (2019). doi:10.1073/pnas.1811878116.
- [19] Wilson, R. C. & Collins, A. G. E. Ten simple rules for the computational modeling of behavioral data. *eLife* **8**, e49547 (2019). doi:10.7554/eLife.49547.
- [20] Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proceedings of the National Academy of Sciences USA* **115**, 2600–2606 (2018). doi:10.1073/pnas.1708274114.
- [21] Supabase. Supabase Queues. Official documentation (accessed 10 October 2026).
- [22] jsPsych. Version 8 documentation: running experiments and plugin API. Experiment execution; core API (accessed 10 October 2026).
- [23] Bridges, D., Pitiot, A., MacAskill, M. R. & Peirce, J. W. The timing mega-study: comparing a range of experiment generators, both lab-based and online. *PeerJ* **8**, e9414 (2020). doi:10.7717/peerj.9414.
- [24] Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science* **11**, 702–712 (2016). doi:10.1177/1745691616658637.
- [25] Silberzahn, R. et al. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science* **1**, 337–356 (2018). doi:10.1177/2515245917747646.
- [26] Kriegeskorte, N. & Douglas, P. K. Cognitive computational neuroscience. *Nature Neuroscience* **21**, 1148–1160 (2018). doi:10.1038/s41593-018-0210-5.
- [27] Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018 (2016). doi:10.1038/sdata.2016.18.
- [28] Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E. & Zou, J. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature* **646**, 716–723 (2025). doi:10.1038/s41586-025-09442-9.
- [29] Alber, S., Chen, B., Sun, E., Isakova, A., Wilk, A. J. & Zou, J. CellVoyager: AI CompBio agent generates new insights by autonomously analyzing biological data. *Nature Methods* **23**, 749–759 (2026). doi:10.1038/s41592-026-03029-6.
- [30] Wang, Z., Danek, B., Yang, Z., Chen, Z. & Sun, J. Making large language models reliable data science programming copilots for biomedical research. *Nature Biomedical Engineering* **10**, 1732–1746 (2026). doi:10.1038/s41551-025-01587-2.
- [31] Chen, Z. et al. ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery. *ICLR* (2025). Conference record; arXiv:2410.05080.
- [32] Gu, K. et al. BLADE: Benchmarking Language Model Agents for Data-Driven Science. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 13936–13971 (2024). doi:10.18653/v1/2024.findings-emnlp.815.
- [33] Asai, A. et al. Synthesizing scientific literature with retrieval-augmented language models. *Nature* **650**, 857–863 (2026). doi:10.1038/s41586-025-10072-4.
- [34] Gorgolewski, K. J. et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data* **3**, 160044 (2016). doi:10.1038/sdata.2016.44.
- [35] Hadidi, N., Fegghi, E., Song, B. H., Blank, I. A. & Kao, J. C. Spurious alignment between large language models and brains can emerge from non-robust methods and overlooked confounds. *Nature Communications* **17**, 5769 (2026). doi:10.1038/s41467-026-72253-7.