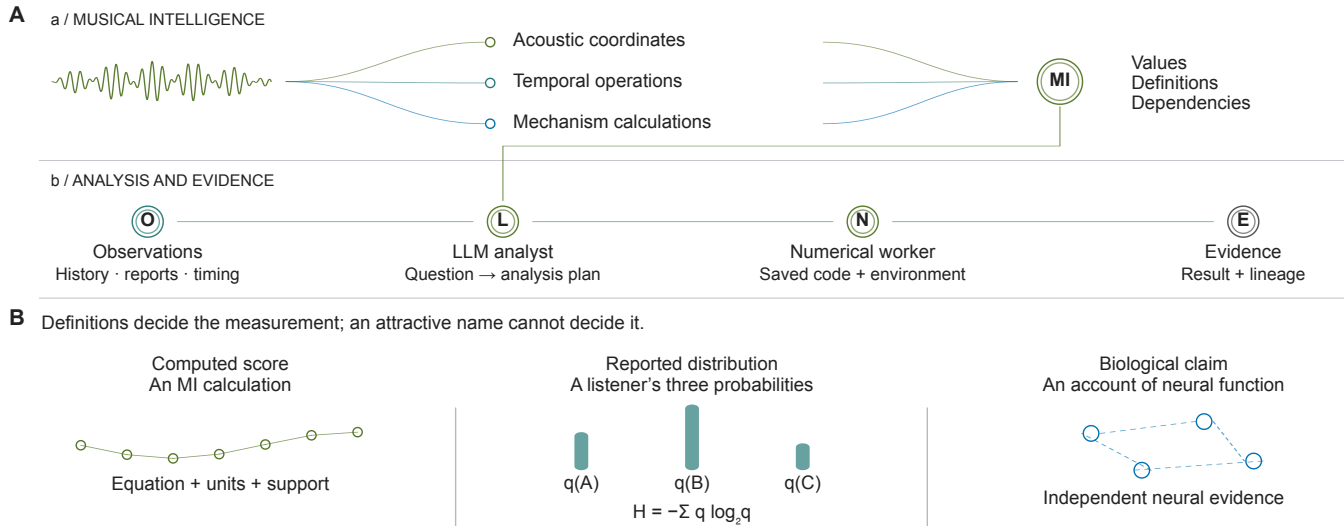


Does model documentation improve the analysis of musical responses?

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

Abstract Can documentation help a language-model analyst draw better conclusions from a computational representation of music? We propose a benchmark using Musical Intelligence (MI), which describes recordings through acoustic and listening-related computations. The main comparison holds numerical arrays and the analyst fixed while adding definitions, temporal support and the record of how quantities were computed. Tasks cover recovery of known relationships, detection of confounds, replication and prediction on held-out data. A successful analysis must produce a supported conclusion from permitted information and numerical results that can be rerun. Matched resources, alternative labels and an expert-written analysis will test competing explanations for any benefit. Musical pleasure is the initial application. This paper describes the framework and planned benchmark; it reports no participant, neural or benchmark results.

Prospective author manuscript. All figures are conceptual or analytic illustrations.



A correspondence is a hypothesis to test.

Pleasure ratings are observations. An MI reward label is not a dopamine measurement.

Fig. 1. Representation, analysis and evidence **A**, MI describes a recording, MI-Lab collects independent observations, and a language model assists analysis planning while numerical tools execute it. The researcher sets the question, access rules and acceptance criteria. **B**, Computed descriptors, reported probabilities and biological interpretations require different definitions and evidence. Curves, bars and circuit-like marks are illustrations, not participant observations or anatomy.

From musical representations to scientific claims

A computational description of music is useful when it helps distinguish explanations for a listener's response. The same pleasure rating may reflect acoustics, expectations, familiarity or their combination. We ask whether documenting how MI quantities are defined and derived helps an analyst compare these explanations. MI supplies the representations, independent observations provide the outcomes, and a language model assists with analysis through saved numerical programs.

The executable architecture[1] contains acoustic coordinates (R^3), requested temporal operations (T^3) and listening-related computations (C^3). These are not necessarily a serial chain: an analysis must record the routes used, including direct acoustic access and fallbacks. The pinned manuscript reports 97 acoustic channels, 84 named mechanisms and 889 mechanism-output dimensions. These describe the implementation, not a set of independent psychological constructs. A study should select a small set justified by its question, with broader searches labelled exploratory and shared inputs accounted for.

An equation specifies a transformation; assigning it psychological meaning also requires construct validation[2]. Analyses may concern transitions, passages, excerpts, performances or people. Contrasting sounds may occur within a single recording, so composition, performance, recording and excerpt identifiers will determine evaluation groups and keep related material together across splits. Engine reconstruction is described in the architecture record.

Pleasure provides the initial application. Uncertainty–surprise effects and individual preference estimates come from different tasks and models[3, 4], and need not imply a single pleasure rule. Sensory surprise and signed value error have different

definitions[5]; neurochemical interpretations require biological assays[6]. Likewise, selected music, logged exposure, recognition and autobiographical recollection are different observations. Recollection would require its own consented assay. Predicting perceived valence or arousal from acoustics[23] does not establish prediction of an individual's experienced pleasure.

Recent tool-using systems provide relevant methodological precedents. The Virtual Lab has been assessed through experimental nanobody design[7]; CellVoyager uses notebooks for biological data analysis[8]; explicit analysis plans have been investigated for biomedical programming assistance[9]. ScienceAgentBench evaluates generated programs and execution[10], BLADE allows multiple defensible analyses[11], and OpenScholar addresses source-grounded synthesis[12]. These studies motivate a music-specific evaluation. In that evaluation, source relevance, numerical correctness and scientific interpretation must each be assessed: a fluent report or a program that runs successfully is insufficient on its own.

Definitions, computational sources and available information

The proposed framework associates each numerical quantity with five groups of metadata: identity and version; units and scale; temporal and population scope; derivation and dependencies; and evidence status. These records allow an analyst to determine what a value means before choosing a model. A value between zero and one is not necessarily a probability, and similarly named channels may measure different things. Replacing probability-defined surprise with an acoustic score, or a rating with a rank, changes the target or loss interpretation. Such changes must be explicit; absent metadata remain unknown.

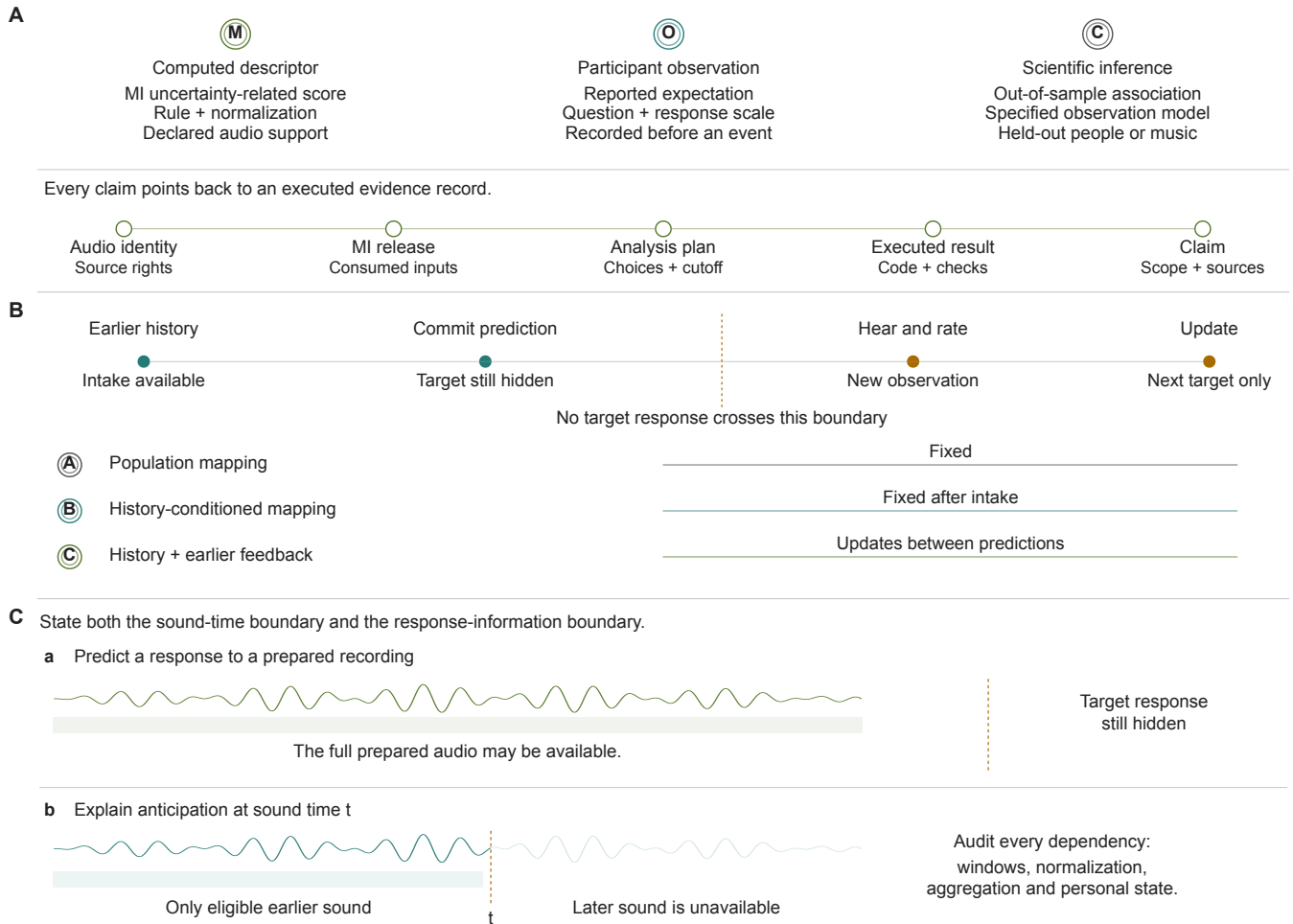


Fig. 2. Information available to an analysis **A**, Definitions, versions and provenance connect quantities to observations and claims. The full record also includes observation releases and exclusions beyond the links shown. **B**, Population, history-only and feedback-updated readouts predict identical targets before those responses are revealed. Their performance ordering is unknown. **C**, Full audio from a known recording may be available for response prediction; a time-local anticipation test must exclude later sound throughout the dependency chain.

Stimuli and observations Stimulus records contain composition, movement and arrangement where applicable, performance, recording and excerpt identifiers, waveform digest, decoding and preprocessing, source permissions, channel count and sampling rate. Observation records add participant pseudonym, session, exposure, trial, wording and scale, response phase, timestamps, completeness and quality flags. A title match alone cannot identify a recording or its permitted derivatives. An identical upload retry preserves one event, whereas repeated listening creates another exposure. Missing responses remain distinct from zero.

MI quantities and personal state Each quantity records its engine, configuration and equation versions, channel identifier, units or range, transformations, normalization, temporal support, availability, consumed inputs, fallbacks and aggregation. Personal-state records add intake version, eligible earlier observations, learner specification, fitted state, seed, update time and digest. The dependency record concerns inputs actually consumed, which may differ from edges drawn in a diagram. Profile changes create new versions. Neither the current target nor later reports are eligible inputs to a state used to predict that target.

Analysis records Each record will identify the question, population, outcome, estimand, predictors, exclusions, transformations, data splits, tuning, multiplicity, uncertainty and failure criteria. Confirmatory choices and exploratory branches will be distinguishable. Execution records will retain code, dependencies, environment, data version, model configuration, prompts, retrieved sources, tool calls, errors, seeds, runtime, cost and outputs. Conclusions will link to estimates and intervals, their scope, supporting results and observations, relevant source passages and review decisions. Citations will distinguish theoretical motivation, equation origin, methodological precedent and direct evaluation.

The analyst will have no write access to evaluation labels or acceptance tests.

Checking dependencies For the executed parents $\text{pa}(v)$ and direct sources $B(v)$ of a value v , define

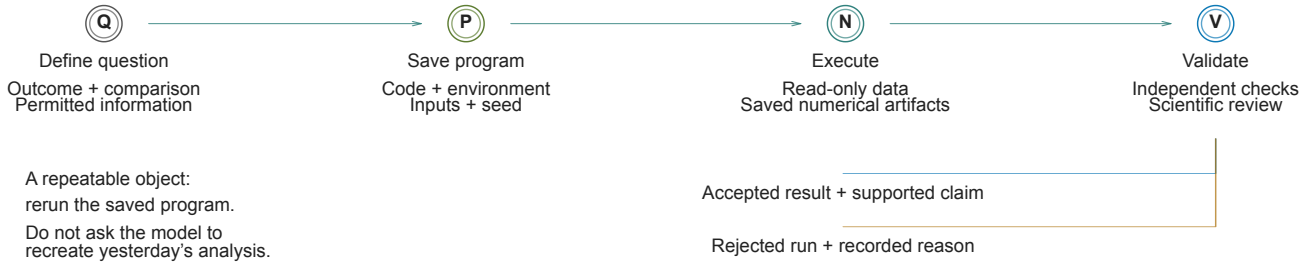
$$S(v) = B(v) \cup \bigcup_{u \in \text{pa}(v)} S(u), \quad S(\hat{y}_k) \subseteq I_k,$$

where I_k is the information permitted for target k . Dependencies include fitted transforms, normalization statistics, personal-state inputs and random streams or generation rules. If all sources and parents are known in an acyclic executed graph, checking them recursively establishes whether the prediction respects its information boundary. This guarantee is conditional on the record being complete. Dynamic reads and caches can introduce untracked inputs, so execution must capture dependencies, restrict access and reject unsupported reads.

Earliest and latest support bounds will propagate conservatively, with unknown support retained. A locally past-labelled operation can still contain future information inherited from an upstream input; storing a feature before a session does not make it causal. Audio-time availability and response-information availability are separate constraints. Hashes establish artifact identity, while access controls are needed to prevent an analyst from reading answers elsewhere.

These checks can identify invalid joins, inconsistent units, information leakage, duplicated exposures and claims unsupported by the recorded evidence. Construct validity still requires separate assessment. FAIR metadata[13] will need interoperability tests, and missing information will limit interpretation. The record covers observable plans, tools, programs and outputs. A proprietary model's internal processes remain inaccessible, so its account of its reasoning cannot replace recorded execution.

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.

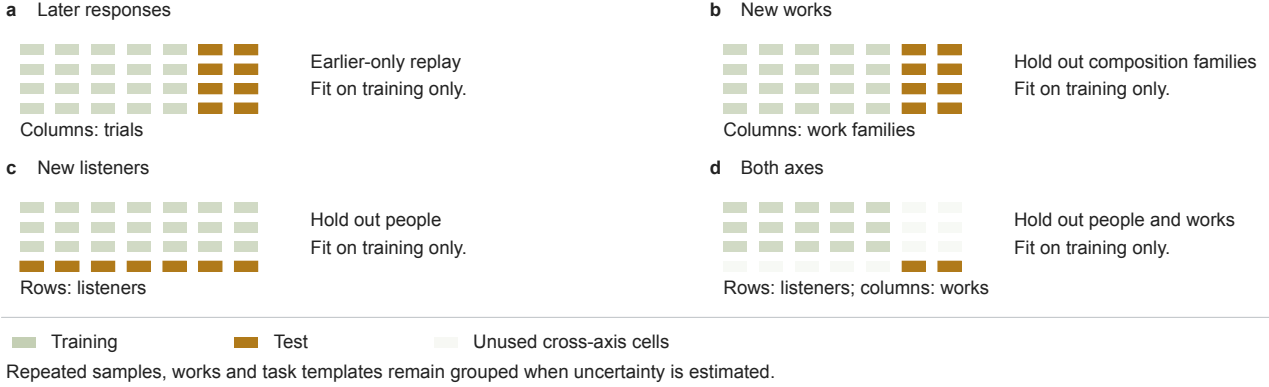


Fig. 3. Execution and generalization **A**, A numerical program is saved, executed and independently checked. Accepted claims and rejection reasons are retained. Reproducibility concerns rerunning that program, rather than generating a replacement. **B**, Holdouts for later responses, new compositions, new people and both new people and compositions address different questions. Unused cells remain unused. The diagrams show conceptual partitions rather than sample sizes. All competitors share outer folds and training-only transformations.

Personalization and the timing of prediction

With frozen MI configuration θ , let $z_s(t) = F_\theta(x_s; t)$. For eligible intake or history h_i , earlier observations $D_{i,<k}$ and context c_k ,

$$w_{i,k} = \mathcal{U}_\phi(h_i, D_{i,<k}), \quad \hat{y}_{i,k} = g_\phi(z_{s_k}, w_{i,k}, c_k).$$

Here, $w_{i,k}$ represents the personal state used by the readout. A language model may help design the update and prediction functions, but the evaluated learner will be saved and fixed before test responses are exposed. Supplying context to an analyst is a separate operation from listener learning or computational personalization.

Three readouts will be compared on identical targets, audio and non-personal covariates: A uses a population model without personal history or feedback; B is conditioned on intake and then fixed; C uses the same intake plus earlier feedback. B–A estimates the contribution of history and C–B that of feedback. A fixed direct-MI baseline requires an explicit target definition; a fitted population mapping is not raw MI. Acoustic counterparts will receive the same history and updating opportunities. Personal-mean, repetition and size-matched swapped-history controls will address scale calibration and generally high rating tendencies.

An online forecast must be committed before the response occurs. Chronological replay can estimate earlier-only prediction, but does not retroactively become an issued forecast. Trusted storage and access restrictions complement timestamps and hashes. Full prepared audio may be eligible for predicting a later response while remaining ineligible for a time-local account of what the listener could have anticipated.

Numerical execution and validation

Analysis workflow Before confirmatory outcomes are available, the researcher will specify the observations, estimand, predictors and their temporal support, validation split, uncertainty procedure and competing explanations[14]. Exploration may revise these choices, with branches and reasons retained. The analyst receives a data dictionary, access and rights constraints, verified source passages and task summaries that identify row counts, aggregation and truncation. Large numerical arrays remain in storage with restricted access.

Python or R programs and validated numerical functions will run in an isolated computing environment with declared dependencies, read-only data and separate outputs for each run. Reported estimates must come from saved results. The interface may use institutional computing without depending on a particular language-model provider. Records will identify the model and available snapshot, configuration, prompts and tools. Internal reasoning is not required as evidence. Since hosted model aliases can change, regenerating code and rerunning saved code will be assessed separately.

Validation will check recording and participant identity, response phase, denominators, units, missingness, exposure history, information availability and expected properties. Known-answer test cases will assess numerical operations, independent calculations will check central statistics where feasible, and positive and null simulations will assess inference. The analyst's run cannot alter its validation tests. Experts will review scientific adequacy; another language model may provide additional review but cannot determine acceptance. Failures and their reasons will be retained.

Observation models and generalization One candidate model is $y_i(t) = a_i + [K_i * f(z_s, w_i, c)](t) + \epsilon_i(t)$, with reporting delay or smoothing K_i and dependent error. Lag, smoothing and model complexity will be chosen on development data. Overall excerpt ratings cannot establish detailed temporal trajectories; slider traces provide temporal information subject to motor delay, scale use, fatigue and sampling limits. Tapping, recognition, wanting, relief and memory require their own measurements[2].

Evaluation splits follow the intended claim: later trials, unseen composition families, unseen people, or both new people and new families. Performances and repeated samples remain grouped. Normalization, feature selection, dimension reduction and hyperparameters will be learned within training data, using the same outer folds for all competitors. Uncertainty will be estimated at participant/composition or source/template level rather than by counting frames. Parameter and model recovery will be assessed[15]. Constant or missing ratings will not be removed merely because a model fits them poorly.

Exploration and confirmation The analysis record will track searched channels, lags, interactions and subgroups. Error control

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.

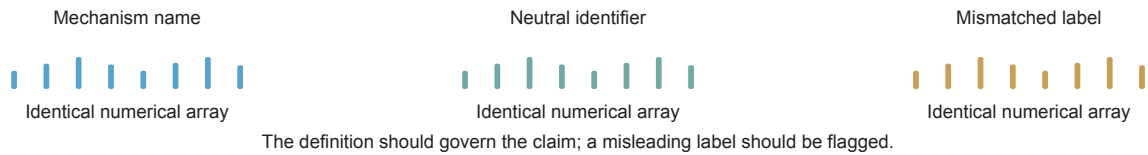
D3 – D1 isolates the evidence contract; it does not establish MI's acoustic superiority.

B

D0	D1	D2	D3
Acoustic	MI values	MI + contract	MI + contract + lineage
● Numeric representation ○ Expanded definitions ○ Dependency / source	● Numeric representation ○ Expanded definitions ○ Dependency / source	● 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. 4. Benchmark comparisons **A**, Representation, documentation and analyst effects require different controls. **B**, D1–D3 use identical MI arrays; every condition includes the minimum units, identifiers, timing and outcome permissions. Definitions and provenance are added in successive conditions. **C**, Neutral or mismatched labels test sensitivity to names, while consumed-edge controls distinguish actual dependencies from drawn connections. These are explicit benchmark manipulations, not changes to participant records.

will be prespecified, and selected discoveries will be evaluated independently. Multiverse analyses can reveal sensitivity[16], while many-analyst findings demonstrate the importance of visible analytic choices[17]; neither supplies repeated independent confirmation. Confirmatory programs will be selected and frozen using development criteria. Newly proposed equations receive new versions and evaluations, leaving failed baselines available.

Execution checks Three properties will be tested. First, canonicalization of an event collection must satisfy $C(E \cup \{e\}) = C(E)$ when e is an identical retry already in E . Changed payloads are conflicts; new playback creates new exposure. Second, dependency checks must detect deliberately introduced future responses in fitted scalars, caches or initial states. Valid cases will include full prepared audio with hidden responses, so blanket rejection of offline prediction cannot pass. Expected decisions will be set before analyst access. Third, saved code, data, environment and random streams must allow an independent rerun. Deterministic operations require exact agreement where supported; numerical and stochastic operations use declared tolerances.

These are proposed tests, not reported operational results. Both validator and analyst failures will be reported. Even a fully reproducible analysis can answer the wrong question, which is why execution checks and scientific review remain separate.

Benchmark design

Information conditions The benchmark separates representation, documentation and analyst effects. D0 provides prespecified acoustic features. D1 provides MI values with a minimal dictionary. D2 adds operational definitions and temporal support. D3 adds executed dependencies, fallbacks and verified source roles. Common

covariates and allowable observations are identical across conditions.

The primary contrast, D3–D1, tests documentation while holding the MI arrays and analyst fixed. D2–D1 and D3–D2 distinguish the contributions of definitions and provenance. D1–D0 compares representation packages, including their dimensionality and read-out, rather than isolating individual mechanisms; equal-dimension and equal-tuning-budget sensitivities will qualify that comparison. A change of model, provider or resources is a separate analyst comparison. Verbosity-matched documentation controls and an expert-authored frozen pipeline will use the same eligible data. Blinded experts will establish task criteria independently, allowing defensible alternatives in open-ended analyses[11].

Tasks and runs The proposed inventory contains 24 development tasks and 120 evaluation tasks: 30 each for known-truth recovery, null or confound diagnosis, replication of accessible analyses, and held-out prediction or discovery. Recovery tasks vary delay, crossed participant/item effects, missingness and planted relationships. Null tasks include irrelevant correlated channels, duplicated compositions, future-dependent preprocessing and unsupported biological claims. Replication tasks require permitted data and executable references. Possible exposure during model training will be recorded; newly generated instances or prospective observations will be reserved for generalization, with private seeds and labels inaccessible to the analyst. The inventory has not yet been collected or executed, and its proposed size requires a precision justification.

Each of four conditions will run three times per task in fresh contexts, yielding 1,440 main runs before controls. Mean success across attempts is the outcome. Timeouts, invalid results and failed repairs count as failures, with infrastructure-outage rules, repair op-

portunities and attempt costs defined equally across conditions. A task passes only if it meets its numerical or empirical target, respects the information boundary, supplies rerunnable artifacts and supports its conclusion without disallowed claims. Qualified or negative conclusions may pass an underidentified or null task; refusing every task cannot.

Numerical tolerances will follow reference calculations and measurement precision. Discoveries will be scored using prespecified recovery or unseen observations, supplemented by expert rubrics for defensible alternatives. Two blinded reviewers will assess claim support and retain their disagreements and resolutions. Language-model-only grading will not determine acceptance.

Primary estimand For binary task success s_{qcr} and its repetition mean $\bar{s}_{qc}$, the documentation effect is

$$\Delta = Q^{-1} \sum_{q=1}^Q (\bar{s}_{q,D3} - \bar{s}_{q,D1}).$$

This estimate describes the sampled task inventory. Paired, stratified resampling will operate on independent source/template families, keeping related tasks and repetitions together. Sensitivity models will include task family and dataset. If independent families are too few, results will remain descriptive. Family definitions, resampling and interval methods, stopping rules and the minimum relevant gain will be preregistered. Development simulations will determine achievable precision and cost and may motivate changes to inventory size. Task-level denominators and intervals will accompany aggregate rates.

Secondary outcomes include numerical error, rerun agreement, unauthorized access, unsupported claims, citation support, expert correction time, latency and total computation cost. An increase in unsupported biological claims will be reported even if completion improves. Representation and model-choice comparisons remain secondary unless separately registered. Additional primary contrasts require familywise error control; exploratory findings require new evaluation. A small benchmark with no observed rare errors cannot establish that their risk has been reduced.

Matched resources and negative controls Conditions will share the base-model snapshot, tools, permissions, libraries, maximum attempts and task-specific budgets. Actual tokens and resources will be reported because equal caps need not imply equal use. Length-matched controls address the opportunity provided by a longer prompt. A smaller or lower-effort analyst will assess model dependence, with exact provider and snapshot recorded at execution.

Neutral and deliberately mismatched labels will preserve values and definitions while testing whether suggestive names influence conclusions. Further controls will remove an unconsumed edge, swap size-matched histories, preserve autocorrelation while breaking target alignment, and insert known nulls. Each control will have a stated expected effect; removing an unused edge should not change the numerical computation. Evidence for the framework would weaken if documentation improved explanations without valid task completion, if gains disappeared under matched resources, or if simpler acoustics performed equally well at lower cost. Benefits limited to a task family will be reported at that scope.

Application to music research

Pleasure and candidate mechanisms The earlier companion protocol[18] specified 20 recognized classical families. Its 12 October integrated revision replaces that requirement with calibration from familiar music, recommendations and unfamiliar music. Explicit reports supply the personal information; neither version reconstructs autobiography or streaming history. Matched predictions can compare B–A and C–B with personal-mean and equally informed acoustic controls. Suitable observations could also compare probability-defined information content with ICEM or PUPF, and resolution-related predictors with roughness, energy and general acoustic change.

These comparisons require precise endpoints. Reported signed value error differs from sensory unexpectedness and RPEM, and algebraically coupled differences cannot establish value learning; independent later behaviour is a more informative target. Mem-

ory claims must be tied to measured exposure, recognition or recollection, while brain prediction requires independent neural observations. ICEM (Information Content Emotion Model), PUPF (Pleasure-Uncertainty Prediction Function), RPEM (Reward Prediction Error in Music), WMED (Working Memory-Entrainment Dissociation) and UDP (Uncertainty-Driven Pleasure) identify implementations. Their equations, temporal support and shared inputs must be inspected before attributing distinct effects. Similar names neither merge UDP with PUPF nor demonstrate their independence.

Consider a musical ending. Events can be selected independently of pleasure, with response interval and motor lag fixed from development data. Equally constrained acoustic-change and context-resolution models can then be compared while accounting for familiarity, repetition and scale use. Strong collinearity may leave the explanations underidentified. A subsequent experiment could hold the ending fixed while varying context, or vary the ending within a controlled context, and measure remaining confounds. Musical validity, rights, participant burden and unintended changes require researcher review before predictions and analyses are committed.

Adaptive selection is an optional extension. It would require records of policy, candidate set, randomization and selection probabilities, together with an analysis of distribution shift. Prepared sessions and offline analysis remain the baseline. A failed direct readout may coexist with useful information elsewhere in the representation, but a post hoc replacement needs an independent test.

Acquisition and access The measurement companion[19] separates released stimuli and schedules, browser exposures and immutable analysis views. The analyst cannot rewrite observations or committed predictions, change collection protocols, coach listeners, supply consent or invent missing responses. Prepared sessions therefore remain usable during analyst-service outages. Audio-context, browser-monotonic and receipt clocks serve distinct roles; timing needs physical qualification[20], and upload retries remain distinct from repeated listening.

Research views exclude identities and contacts, although pseudonyms and small-group summaries may still reveal individuals. Tool access will be limited to the task, with aggregation and suppression recorded. Hosted-language-model data flow must be addressed in consent and institutional review before transmission. Institutional numerical workers may return approved summaries, without implying that every summary is anonymous. Recollection data require a distinct measurement and handling purpose, and the proposed analyses do not generate participant diagnoses.

Papers, participant text and metadata will enter the analysis as source material, without authority to change access permissions. Computation will have bounded resources, separate outputs and restricted network access. Evaluation outcomes will remain unavailable until predictions are saved. Questions, definitions, plans, scoring rules, code and permitted derivatives will be released; identifiable records and restricted audio will require controlled access or reproducible test data. Reproducibility reports will state what is available publicly, to approved reviewers and internally, including architecture-related restrictions[1]. File hashes cannot replace the underlying files or an independent rerun. Access and reuse conditions will follow FAIR principles[13].

Neural studies EEG and MEG applications require stimulus alignment, preprocessing, sensor or source definitions and temporal kernels. fMRI applications require haemodynamic assumptions, run structure, motion controls and region selection. BIDS[21] can support consistent identity and metadata. Encoding comparisons additionally need held-out signals, matched acoustic, timing and nuisance baselines, and shared folds and tuning. Test outcomes cannot be used to select regions, time windows or models.

Work on apparent language-model–brain alignment[22] motivates controls for shuffled splits, autocorrelation, recording position, rate or envelope and leakage; it does not evaluate MI. The most informative manipulations make competing explanations disagree. Code interventions address computations, randomized sound or context addresses assigned behavioural effects, and neural mechanism claims require evidence at that biological level[24].

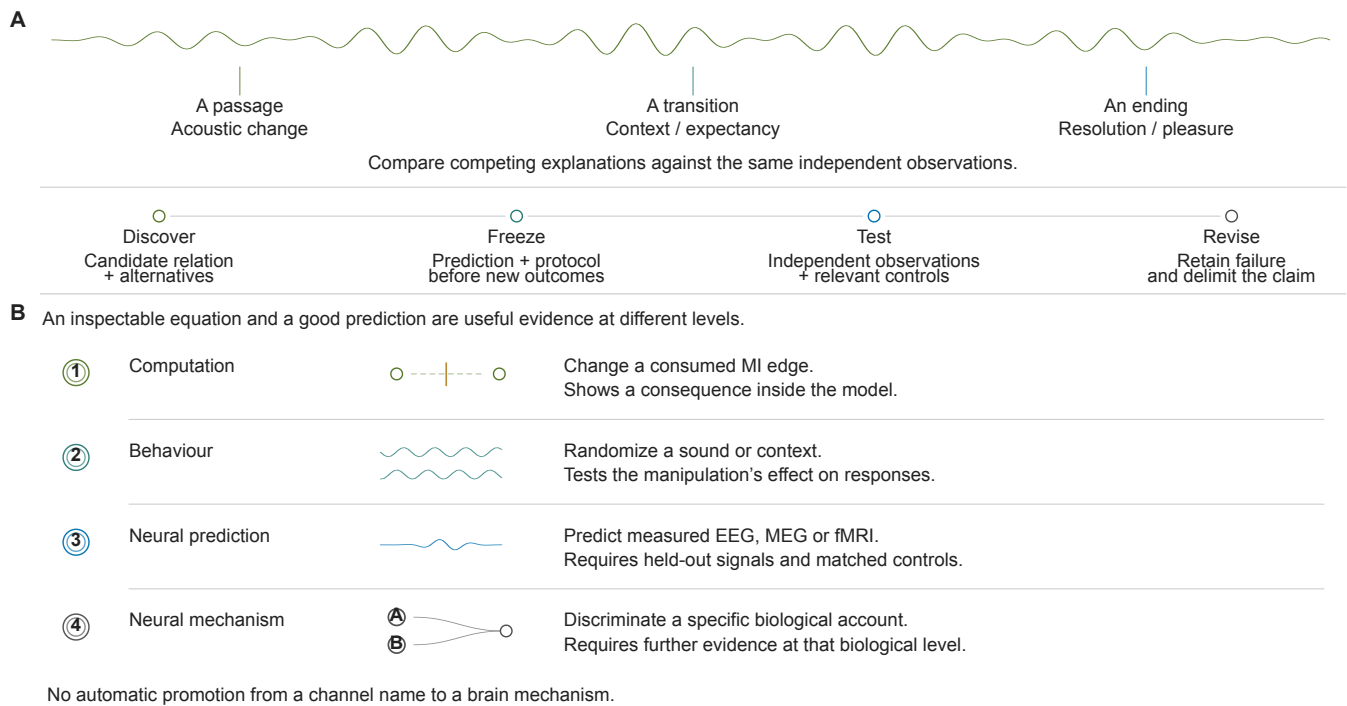


Fig. 5. Discovery and levels of evidence **A**, Competing passage-level explanations lead to a new test whose predictions are fixed before evaluation. Changing code and manipulating a listener's experience test different claims. **B**, Computational results, randomized behaviour, held-out neural prediction and specific biological mechanisms require evidence appropriate to each claim. The diagram identifies these distinctions rather than reporting completed validation.

Interpretation and limitations

MI can compute a quantity reliably while its proposed memory or reward interpretation remains unsupported. It may also predict responses without distinguishing among mechanisms. The benchmark tests whether access to equations, assumptions and shared inputs helps an analyst recognize these limits and reach better conclusions. Reproducible numerical programs do not require access to language-model internals, while reconstructing the engine requires its own materials. Results from classical-music listeners would need further tests across styles, cultures, ages and listening contexts.

Benchmark limitations include fallible expert judgment, possible training exposure to public replication tasks, incomplete synthetic nulls and dependence among datasets. Many repeated runs cannot compensate for few independent task families. Provider changes, costs and stochastic analyst variation also differ from the reproducibility of frozen numerical code. The inventory, reference calculations, rubrics, family assignments, budgets, scoring and development decisions will be fixed and made auditable before evaluation, while private labels remain inaccessible. Permitted artifacts and failures will be released afterwards.

A failed check can narrow or invalidate a claim, or motivate a new comparison; it does not automatically justify deleting observations. Underidentification should be attributed to shared predictors, uncertain measurements or missing contrasts where supported. Failed generalization and unavailable reconstruction records will be reported as such.

The first step is an independent rerun on permitted data, including a deliberate failure case, followed by a blinded comparison of numerical inputs with and without documentation. Prospective personalization and discovery tests, and later neural studies, depend on the required measurements and expertise. A lack of measurable documentation benefit would provide grounds to revise or discontinue the framework.

Status and disclosures This paper reports no new participant or neural data, benchmark findings, deployed analysis service or completed preregistration. Each proposed study requires an analysis plan, sample justification and institutional approval. Editable text, vector figures and review records accompany the manuscript; engine access follows the architecture record. Amaç Erdem develops MI and proposes MI-Lab. Funding and competing interests will be confirmed before submission. Generative AI assisted literature retrieval, drafting, figure programming and document checks; no par-

ticipant records were supplied. This editorial assistance does not evaluate the proposed analyst. The author remains responsible for reviewing the content and disclosure before submission.

References

- [1] Erdem, A. Musical Intelligence: A Glass-Box Architecture for Computational Models of Music Listening. Author manuscript, 8 October 2026. Architecture record. Implementation and evidence claims are those of the cited author manuscript.
- [2] 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.
- [3] 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.
- [4] Mas-Herrero, E. & Marco-Pallarés, J. Predictive processes shape individual musical preferences. *Proceedings of the National Academy of Sciences USA* **122**, e2500494122 (2025). doi:10.1073/pnas.2500494122.
- [5] Gold, B. P., Mas-Herrero, E., Zeighami, Y., Benovoy, M., Dagher, A. & Zatorre, R. J. 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.
- [6] Salimpoor, V. N., Benovoy, M., Larcher, K., Dagher, A. & Zatorre, R. J. Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. *Nature Neuroscience* **14**, 257–262 (2011). doi:10.1038/nn.2726.
- [7] 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.
- [8] 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.
- [9] 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.

- [10] Chen, Z. et al. ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery. *ICLR* (2025). Conference record; arXiv:2410.05080.
- [11] 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.
- [12] Asai, A. et al. Synthesizing scientific literature with retrieval-augmented language models. *Nature* **650**, 857–863 (2026). doi:10.1038/s41586-025-10072-4.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] Erdem, A. Individual expectations and adaptive prediction of musical pleasure. Author manuscript, journal preparation revision 1, 10 October 2026. Distributed with this edition; not a published result.
- [19] Erdem, A. Measuring musical experience across time: the MI-Lab framework. Author manuscript, journal preparation revision 1, 10 October 2026. Distributed with this edition; not a published result.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] Gu, Y., Shao, C., Li, J. & Fan, Y. Interpretable deep learning reveals distinct spectral and temporal drivers of perceived musical emotion. *Scientific Reports* **16**, 4205 (2026). doi:10.1038/s41598-025-34238-2.
- [24] Kriegeskorte, N. & Douglas, P. K. Cognitive computational neuroscience. *Nature Neuroscience* **21**, 1148–1160 (2018). doi:10.1038/s41593-018-0210-5.