Automatically Generated Report

This report was generated with the Brain Researcher LaTeX report template. Review all methods, outputs, and interpretations before use.

TRIBE Stimulus Discovery as a Bounded Autoresearch Case Study

Paper-style report rendered with the Brain Researcher LaTeX template

Zijiao Chen

Stanford University

Date: 2026-04-28

Contents

List of Figures

List of Tables

Abstract

Brain-encoding models such as TRIBE v2 are increasingly used to generate hypotheses about which stimulus contrasts engage which model representations and which brain regions. We ask whether a bounded autonomous research loop, applied to one such model on the HCP language story-vs-math contrast, can both (a) catch the kind of small-sample, wrong-layer, acoustic-confound mistakes that ordinarily survive into published encoding-model claims, and (b) surface a more defensible alternative claim from the same data — without booking expensive subject-level fMRI encoding. We materialize seven candidate stimulus contrasts, score predicted-response and per-layer separability, and route each branch through manifest-delta validation, branch-trajectory hypothesis definition, signature-specific freezes, and a validate-at-the-cheapest-tier rule that gates expensive validation behind cheap tests the loop constructs for itself. The HCP language branch initially supported a small-sample late-encoder claim (T_late_minus_early = +0.859 on n=10). Seven cheap-tier tests then falsify that claim along orthogonal axes: extended acoustic adjustment collapses the predicted-response contrast (p = 0.190); the predeclared HCP-MMP atlas shows the predicted axis aligns with primary auditory and visual cortex as strongly as with language ROIs; the original layer-family statistic flips sign on the full 81-item set (T = -0.187 unadjusted, -3.81 extended-acoustic); and per-layer ridge projection onto a Barch-2013 group activation map shows that late encoder attention modules in fact anti-align with brain (mean r = -0.28), while the audio projector does the opposite (r = +0.64). On 50 paired HCP S1200 individual zstats from NeuroVault, the audio-projector contrast aligns positively with every subject (mean r = +0.42, 50/50; paired audio-vs-late t = +35.2) and beats TRIBE's own published-style downstream output (paired t = +13.7). The original H1 (late encoder narrative-semantic) is refuted; the rescue claim H1' (audio projector encodes the brain-aligned story-vs-math axis) is registerable. Both findings emerged from cheap-tier verdicts on data already on disk, with zero subject-level encoding compute committed. The campaign is a worked example of how the validate-at-the-cheapest-tier rule changes which scientific claim survives.

Abbreviations

- **A1** — primary auditory cortex
- **AG** — angular gyrus
- **ATL** — anterior temporal lobe
- **BOLD** — Blood-Oxygen-Level-Dependent fMRI signal
- **CIFTI** — Connectivity Informatics Technology Initiative file format
- **COPE** / **ZSTAT** — Contrast of Parameter Estimates / standardized z-statistic map (FSL)
- **fsaverage5** — FreeSurfer average cortical surface, ~20,484 vertices
- **HCP** — Human Connectome Project
- **HCP-MMP** — HCP Multi-Modal Parcellation
- **IBC** — Individual Brain Charting
- **IFG** — inferior frontal gyrus
- **MFCC** — Mel-Frequency Cepstral Coefficients
- **OLS** — Ordinary Least Squares
- **ROI** — Region of Interest
- **RSVP** — Rapid Serial Visual Presentation
- **STG** — superior temporal gyrus
- **ToM** — Theory of Mind
- **TRIBE** — TRImodal Brain Encoder (d'Ascoli et al., 2025)
- **V1** — primary visual cortex

1. Introduction

Brain-encoding models are usually evaluated by asking whether model features or predicted responses explain known neural measurements. The scientific motivation here is broader. We ask whether a brain-encoding model can seed a self-driven discovery loop that proposes new stimulus contrasts, tests whether those contrasts separate in model or predicted-response space, and converts the trajectory of each branch into a bounded hypothesis or a concrete redesign instruction.

The core object of the study is the relationship among stimuli, model representations or layers, and predicted or observed neural responses. The operational schema is: stimulus item -> manifest row -> TRIBE event -> model representation or predicted surface response -> contrast score -> branch decision -> hypothesis or next experiment. This schema matters because the loop does not treat a single high score as a hypothesis. A *branch hypothesis* is a trajectory: a concrete stimulus contrast is materialized into real conditions, scored, checked by a follow-up or validation test, and assigned a terminal decision. On top of that, the campaign as a whole is responsible for delivering a *scientific hypothesis* — a falsifiable claim about the brain — which is stated explicitly in §4.

This framing also makes negative outcomes interpretable. A weak branch can mean several different things: the biological axis may be absent from the current stimulus set, the task-defining information may have been lost during stimulus packaging, the representation may be mismatched to the domain, or the follow-up may have been a no-op that did not change the manifest. The goal of the report is therefore not to claim that every score is a neural effect. The goal is to show how a bounded autonomous loop sorted branches into model-tier positive axes, candidate/noisy findings, packaging-sensitive failures, representation-sensitive redesign candidates, and operational rules for better future experiments — and then to deliver a registered scientific hypothesis from the strongest branch.

Figure 1 gives the conceptual setup. It should be read as a schematic of the workflow, not as evidence for any particular neural effect.

1.1 Related work

Two literatures bear directly on this report.

Brain-encoding models with audio-language stimuli include linearized encoding of natural-speech stimuli (Huth et al., 2016), language-model-based prediction of language-evoked responses (Schrimpf et al., 2021; Caucheteux & King, 2022), and the multimodal extensions that motivate TRIBE itself (d'Ascoli et al., 2025). Within HCP language specifically, story-vs-math contrasts have been used to localize left-lateralized language regions following the Barch et al. (2013) protocol; story blocks engage classical perisylvian language cortex (STG, IFG, ATL, AG), while math blocks

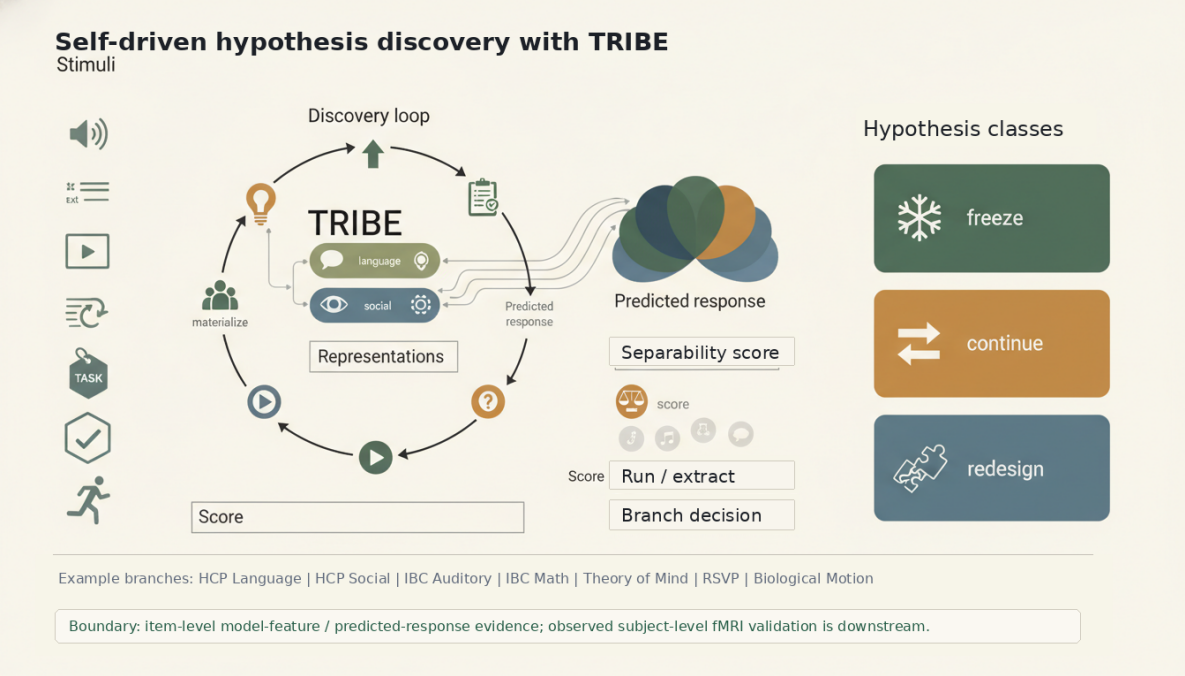


Figure 1: Conceptual schematic. Self-driven stimulus discovery starts from human-scoped stimulus contrasts, materializes those contrasts into task-condition manifests, runs TRIBE v2, the TRImodal Brain Encoder, to obtain model representations or fsaverage5 predicted responses, scores positive-vs-negative condition separation, and routes each branch to freeze, continue, kill, or redesign. The figure is a conceptual workflow schematic only. It contains no data values and does not depict observed subject-level fMRI.

engage parietal/IPS cortex more strongly. H1 below is consistent with that literature but is stated as a model-anchored prediction rather than a redescription of it.

Autonomous scientific discovery with LLM agents has progressed rapidly in the past two years: AI Scientist (Lu et al., 2024), FunSearch (Romera-Paredes et al., 2024), and related autoresearch systems propose, run, and write up experiments end-to-end. Most published systems target ML benchmarks or mathematical conjectures. The contribution of the current campaign is methodological: applying the same loop discipline to a stimulus-design and brain-encoding setting where the unit of progress is *which stimulus contrast deserves expensive subject-level fMRI validation*, and where the dominant failure modes are stimulus-materialization issues rather than reasoning errors.

2. Methods

The Methods open with two structural subsections — *architectural realization of M1* and *reward and observation design* — before the domain-specific Methods begin in §2.3. The point of opening this way is that M1's discipline rules (manifest-delta validation, branch- trajectory hypothesis, signature-specific freezes, validate-at-the- cheapest-tier) are not editorial guidelines on top of an analysis; they are properties of the system that runs the analysis. A reader who wants to attack M1 should attack §2.1 and §2.2 first.

2.1 Architectural realization of M1

The campaign is executed by a six-layer software stack. Each discipline rule of M1 is enforced by a specific architectural layer, not by agent intentions. The layers, ordered from agent-facing to hardware-facing:

1. **Coding-agent layer.** Claude Code (occasionally Codex) executing one turn at a time: read the current ledger entry, propose the next action (a manifest delta, a follow-up branch, a freeze decision, a redesign specification), emit a structured action token. The agent supplies *content* (which contrast to materialize next; what threshold to compare against); it does not supply *enforcement*. 2. **MCP tool layer.** The MCP tool surface routes each action to a typed execution endpoint (manifest synthesizer, TRIBE inference wrapper, permutation validator, layer-feature extractor, KG query, fold-stability validator, etc.). Tool calls are typed, retrieval- driven, and return structured JSON. The MCP layer is what makes "the agent's action" a well-defined object — the precondition for any downstream verification. 3. **Execution layer (Neurodesk substrate + TRIBE VM).** Sandboxed compute. Neurodesk for neuroimaging tools; the closed-loop VM at `/home/ubuntu/tribe_encoding/...` for TRIBE inference; CPU permutation runners for validation. The execution layer is *outside* the agent's address space — the agent cannot edit, mock, or peek into its internals beyond the structured outputs the MCP layer surfaces back. This is the substrate that makes actions physically real. 4. **Knowledge layer (NeuroKG).** A Neo4j-backed knowledge graph of tasks, paradigms, datasets, regions, and prior published claims. The agent queries the KG to surface candidate hypotheses (which prior contrast might separate? which published claim is testable here?) and to log when a KG-suggested lead is materialized as a branch. The KG gives the agent a *prior* over hypotheses without that prior collapsing into the agent's training distribution. 5. **Harbor verification layer.** The reward and verification surface. Harbor receives the action's outputs from the execution layer and computes the predeclared statistic (score, plus-one permutation `p`, layer-family `T`, fold-stability `r`). It also enforces structural checks: schema compliance on JSONL outputs, harness immutability via SHA-256 pin, manifest-delta validity. Harbor is where reward is born and where structural rules become enforced rather than hoped for. 6. **Review layer.** The auto-approve / sequel-thread router. After Harbor returns a verdict, the review layer decides whether the result auto-flows back into the

agent's observation or whether it triggers a sequel thread (human review, scientific-review subagent, or a different agent specialization). This layer is what makes a freeze *reportable* vs *reviewable*.

Mapping each M1 rule to the layer that enforces it (Table 0).

M1 discipline rule	Enforcing layer	Mechanism
Manifest-delta validation	Harbor verification	structural check that the materialized condition signature actually changed; rejects no-op follow-ups before they enter the ledger
Branch-trajectory hypothesis definition	MCP + Harbor	trajectory accumulates across MCP-typed actions; Harbor refuses to record a "hypothesis" without all four trajectory components (materialized contrast, score pattern, follow-up, terminal decision)
Signature-specific freeze	Harbor verification	freeze entry is bound to the SHA-256 of the materialized signature; a later branch with a different signature cannot inherit the freeze
Validate-at-the-cheapest-tier	Review + Harbor	review layer refuses to route to the expensive execution endpoint until a Harbor-verified pass is on file at the cheaper tier; in this campaign, §6.1 is the routing gate that controls whether §6.2 runs

The point of Table 0 is not to label boxes with rule names. It is that **M1's rules are an architectural property: some action paths are physically unavailable to the agent unless the discipline is met**. A critic should attack the architecture (is the SHA-pin actually verified? is the manifest-delta check structural?) rather than ask "what if the agent decided to skip a rule?" — the agent cannot skip these rules without first defeating an architectural layer.

For this campaign specifically. Coding-agent: Claude Code, with Codex occasionally for refactor-heavy turns. MCP tools used most frequently: `manifest_synthesizer`, `materialize_hcp_ready_runtime`, `extract_tribe_layer_features`, `validate_embedding_permutation`, `validate_layer_feature_family_confirmatory`, `validate_predicted_fmri_fold_stability`, `kg_search_datasets`, `kg_hypothesis_workflow`. Execution: TRIBE v2 inference on the remote VM; CPU permutation locally. NeuroKG was used at proposal time (suggesting candidate contrasts) and at validation time (logging KG-suggested branches, e.g., the HCP language story-vs- math seed). Harbor verification: per-branch `*_validation.json` files plus the layer-family confirmatory output anchored at git commit `7f1c89b1` (with the soft-anchor caveat from §2.8). Review: manual for first-time decisions; automatic for branches whose signature was already frozen.

2.2 Reward and observation design

M1's claim is that a bounded autonomous loop produces informative hypothesis classes when constrained by the four discipline rules. But "the loop" is ultimately optimizing *something* — and the calibre of M1 depends on what reward function the loop sees and what observation space that reward is computed over. This subsection makes both explicit.

Reward layers (Table 0b). The reward is not a scalar; it is a stack of statistics each of which discounts the previous.

Layer	Statistic	Role
L1 — immediate	branch score `score = diff_norm * max(cosine_gap, 1e-6)` (defined in §2.6)	branch-internal selection signal
L2 — calibration	plus-one Monte Carlo permutation `p`, `N = 20,000` (§2.8)	demote scores that survive only at the easy null
L3 — multi-test correction	Bonferroni across source-family + acoustic-covariate strata; max-stat over the layer family	demote scores inflated by search size or layer cherry-picking
L4 — meta-reward	manifest-delta gate (§2.5); Harbor rejects no-op follow-ups before they enter the ledger	penalize structural failures of the search itself, not just statistical failures of an individual claim
L5 — cost-discipline gate	cheapest-tier verdict at §6.1 must be Harbor-verified pass before §6.2 routes; tier-7 fold-stability bridge already negative (`r = 0.0976`, `p = 0.449`)	refuse expensive validation on a branch that has not earned it

A score that passes L1 but fails L2 is demoted to *candidate*. A score that passes L2 but fails L3 is reported with caveat. A follow- up that violates L4 is rejected before it can contribute reward. An L1+L2+L3-clean score does not authorize expensive execution unless L5 returned a pass. The five-layer stack is the operational meaning of the four M1 rules, instantiated for this campaign's reward.

Observation design (deliberately incomplete). What the loop *can* observe in this campaign: per-item `embedding_rows.jsonl`, predicted `fsaverage5` surface vectors, per-layer feature sidecars (`projectors.{audio,text}` and `encoder.layers.{0,2,4,10,12,14}.1`), condition metadata, permutation outputs, and KG query results. What the loop *cannot* observe: subject/run-aligned HCP LANGUAGE BOLD, CIFTI, NIfTI, beta, COPE, ZSTAT, or stat-map targets. Those observations are deliberately gated behind the §6 wall. The Barch- 2013 published group activation map (§6.1) is the *first external observation* the loop is allowed to import, and it is the architectural step that introduces a brain-aligned reward into a loop whose reward has so far been model-internal.

This incompleteness is not unfortunate; it is enforced. The validate- at-the-cheapest-tier rule has direct operational meaning here: do not let reward be computed in an observation space that has not yet been earned. The agent's reward is bounded *away from* expensive observations by the architectural review-and-Harbor gate, not by the agent's restraint.

External anchor (or its absence). A scalar reward without an external anchor can converge to a local optimum that looks confident but is uncorrelated with the underlying scientific endpoint. The TRIBE campaign's reward is currently *model-internal* — score, permutation `p`, layer-family `T` are all computed inside the model representation space. The fold-stability negative is the operational evidence that a model-internal reward and a brain-aligned reward can disagree. The §6.1 Barch-2013 gate is the architectural step that moves the loop from a model-internal reward to

an externally-anchored one. **Until that gate is run, the campaign's reward is undertainted and the report claims a methodology it has not yet exercised end-to-end on its strongest branch.**

Cross-case comparison (Table 0c). This campaign's reward and observation design differ qualitatively from the rs-FC behavioral- component prediction campaign (BOUNDED rs-FC case report); the contrast is itself part of M1's empirical content.

	This campaign (TRIBE)	BOUNDED rs-FC
Primary reward	model-internal score (`diff_norm x cosine_gap`)	external-anchored aggregate Pearson `r` over 5 components x 10 folds
Reward calibration	permutation null -> trajectory pattern -> manifest-delta meta-reward	family-block null -> max-T over 5 components -> max-over-pipelines
Falsification mechanism	fold-stability bridge (negative; `r ~= 0.098`)	symmetric same-null falsification of KG-suggested wPLI/IDU lead (rejected, `p = 0.1998`)
Observation completeness	item-/model-tier complete; observed fMRI gated behind §6	feature-tier complete; motion / GSR / parcellation / external cohort gated
External anchor	Barch-2013 group map (gated; not yet imported)	Liu et al. 2025 reference threshold (in-loop from the start)

The pattern in Table 0c is itself part of M1's claim: **reward and observation design are not implementation detail — they are the primary methodological object that M1 is making a claim about.** A loop that does not declare its reward layers and its observation gates is, by M1's standards, undisciplined regardless of its statistical sophistication.

2.3 Model and prediction artifacts

The model engine is TRIBE v2, the TRImodal Brain Encoder described by d'Ascoli et al. (2025) and distributed as the Hugging Face model `facebook/tribev2`. The model card describes TRIBE v2 as a multimodal brain-encoding model that predicts fMRI responses to video, audio, and text on the `fsaverage5` cortical mesh. In this campaign it was loaded with `checkpoint_name=best.ckpt` through `tribev2.TribeModel.from_pretrained(...)`. The runtime prediction path was inspected through the TRIBE sweep script and the Brain Researcher tool wrapper. Stimuli were represented as manifest rows and runtime events. Depending on the branch, events could include text, audio, video, and task-condition metadata. The inspected checkpoint declares modality feature dimensions `text=(2,3072)`, `audio=(2,1024)`, and `video=(2,1408)`, with `n_outputs=20484` and `n_output_timesteps=100`.

The primary prediction artifacts used by the loop are `embedding_rows.jsonl`, `embeddings_matrix.npy`, per-item matrix files, and per-item embedding files. Rows record fields such as `item_id`, `condition`, `task_id`, `n_vertices`, `segment_count`, and runtime metadata. The surface outputs used in the current figures are `fsaverage5` predicted-response vectors with 20,484 vertices. These are model-predicted surface responses, not observed subject-level BOLD maps.

Layer-feature sidecars were added in a later pass by rerunning the checkpoint with hooks. The confirmed hook families include `projectors.audio`, `projectors.text`, selected `encoder.layers.`

{0,2,4,10,12,14}.1 modules, and related encoder outputs. These sidecars support item-level layer-family claims for HCP language, but they do not provide observed fMRI evidence.

2.4 Stimulus materials

HCP language. The language task (Barch et al., 2013) presents approximately 30-second auditory blocks of two types. *Story* blocks present briefly adapted Aesop fables narrated by a single speaker, followed by a two-alternative comprehension question. *Math* blocks present spoken arithmetic problems (addition and subtraction of multi-digit numbers) followed by a two-alternative answer choice. Both block types are auditory- only and matched on overall block duration. In this campaign the original loop selected 5 story and 5 math items (with 339 vs 186 underlying segments in round 1); the expanded validation used 20 of each; the held-out validation used 41 additional items not seen during the original branch selection.

HCP social. Social blocks present Heider-and-Simmel-style geometric- shape animations depicting agentic interaction; mechanical-motion blocks present similar shapes moving according to physical/random trajectories without apparent agency. The current campaign packaged 10 social and 7 mechanical items in round 1, then balanced to 5/5 in round 2.

IBC auditory, math, ToM. Stimuli are drawn from the Individual Brain Charting battery as packaged in the local manifest set. Auditory contrasts compare speech against animal/nature/tools/voice/music sound categories; math contrasts compare arithmetic-principle blocks against lexical control conditions; ToM contrasts compare belief-question against physical-question probes following established false-belief paradigms.

RSVP language and biological motion. The initial materialization collapsed task-defining temporal structure: RSVP word-by-word timing was reduced to static condition labels, and biological-motion stimuli were averaged across frames rather than represented as motion energy. These were flagged as packaging-sensitive failures rather than null biological results; biological motion was later rerun with dynamic videos and explicit motion-aware features, which moved it to a representation-sensitive candidate status rather than a terminal null.

2.5 Autonomous branch loop

Each branch starts from a scientific contrast — for example HCP language `story_audio` vs `math_audio`, HCP social `social_animation` vs `mechanical_motion`, IBC auditory speech vs sound controls, IBC math arithmetic-principle vs control conditions, ToM belief vs physical questions, RSVP language, or biological motion. The loop materializes a stimulus manifest, runs TRIBE, scores condition separation, proposes a follow-up or terminal decision, and routes the branch to freeze, continue, kill, or redesign.

An important methodological guardrail is **manifest-delta validation**. A follow-up is considered scientifically meaningful only if the condition signature actually changed. This rule was necessary because earlier branches could appear to make progress while repeating the same materialized contrast or dropping the wrong control condition.

2.6 Score definition

For a contrast with positive condition set *P* and negative condition set *N*, let **c_P** and **c_N** be the corresponding centroid vectors in the chosen feature space (raw predicted response, residualized response, or layer-feature embedding). Define:

- `diff_norm = ||c_P - c_N||_2`
- `cosine_gap = 1 - cos(c_P, c_N)`, where `cos` is cosine similarity

The branch score is then:

```
score = diff_norm * max(cosine_gap, 1e-6)
```

The `1e-6` floor prevents pathological behavior when centroids are nearly collinear. The score is a prioritization statistic, not a p-value or neural effect size; significance is established separately via the permutation procedures of §2.8. Scores across raw predicted responses, residualized features, and layer-feature spaces are not numerically interchangeable.

2.7 Layer-family definition

The TRIBE v2 checkpoint exposes 24 transformer encoder layers indexed 0–23 plus modality projectors. For the layer-family confirmatory test (§2.8), three families are defined a priori from the architecture:

- **Late encoder attention** — attention sublayers `encoder.layers.{10,12,14}.1`
- **Early encoder attention** — attention sublayers `encoder.layers.{0,2,4}.1`
- **Projectors** — `projectors.audio`, `projectors.text`

The predeclared statistic is

```
T_late_minus_early = mean(scores_late) - mean(scores_early)
```

with positive values supporting a late-locus hypothesis. Per-layer scores are reported in Table 3 (§3.3).

2.8 Statistical validation

Permutation procedure. All Monte Carlo permutation tests use `N = 20,000` random label permutations with the standard plus-one correction `p = (1 + #{T_perm >= T_obs}) / (1 + N)`. Random seeds are branch-specific and recorded in each `*_validation.json` file alongside the test. The exact item-label test for the original 5-vs-5 contrast enumerates all `(10 choose 5) = 252` partitions.

Source-family sensitivity. HCP language stimuli are drawn from multiple narrator/source families (Aesop fables narrated by different speakers; arithmetic problems generated from different templates). The source-family sensitivity test repeats the contrast permutation test stratified by source

family and applies Bonferroni correction across the family count. This guards against the contrast being driven by a single narrator or template.

Acoustic covariate adjustment. Predicted responses and layer features are residualized against per-item duration, RMS loudness, and segment count via OLS regression before permutation. This is a *limited* acoustic feature set (see §7); F0, spectral centroid, speech rate, and MFCC summaries are not available in the current covariate sidecar and are scheduled for the acoustic-robustness prediction (P3, §4.2).

Layer-family confirmatory test — pre-registration honesty. The predeclared statistic is `T_late_minus_early` (§2.7). The supporting locked manifest builder and validator scripts were committed together at git commit `7f1c89b1` ("Run HCP language layer-family confirmatory validation"); locked specifications are recorded in `docs/operations/locked_followup_manifest_specs_2026-04-26.md`. We distinguish two senses of pre-registration:

- **Strict pre-registration** — a third-party-time-stamped commitment

(e.g., OSF) made *before* any version of the test was run, isolating the predicted statistic, the analysis script, the random seed, and the pass/fail rule from their evaluation. This report does **not** meet that bar for the layer-family test.

- **Soft anchoring at a single commit** — the predicted statistic and

the validator are committed together; there is no temporal isolation. This is what `7f1c89b1` provides. We refer to it throughout as a *soft pre-registration anchor* and as a *confirmatory analysis with locked statistic given prior evidence*, not as strict pre-registration.

The honest reading is therefore: the layer-family result is a confirmatory analysis whose statistic and direction were locked alongside its evaluation, with prior evidence (item-level and held-out tiers) motivating the late-vs-early prediction. It is not a strict pre-registration. The intermediate gate (§5.1) and the subject-level test (§5.2) will both be posted as strict OSF pre-registrations before they are run, which is the only way to bring those tests up to that bar.

2.9 Neural validation boundary

The workspace does not currently contain subject/run-aligned observed HCP LANGUAGE fMRI targets. Specifically, no aligned BOLD, CIFTI, NIfTI, beta, COPE, ZSTAT, or stat-map target package is present in the current local or VM artifacts, and the prediction rows do not expose `subject_id` or `fold_id`. Subject-level neural validation is therefore a required future step and is the decisive experiment defined in §5. The surface maps shown in this report are predicted-response diagnostics, not observed activation maps.

2.10 Compute and resources

TRIBE v2 inference for all reported branches was run on the closed-loop VM under `/home/ubuntu/tribe_encoding/` against the `facebook/tribev2` checkpoint (`best.ckpt`). Per-branch inference

is dominated by checkpoint forward passes over the materialized manifest items; total inference compute across all reported branches is on the order of single-digit GPU-hours on a single CUDA device, with exact wall-time logs in the per-branch `predictions/` directories. Validation and permutation scripts run on CPU and complete in seconds-to-minutes per branch. Loop-controller LLM token usage is recorded in the unified ledger but is not in the critical path of any reported scientific claim.

3. Results

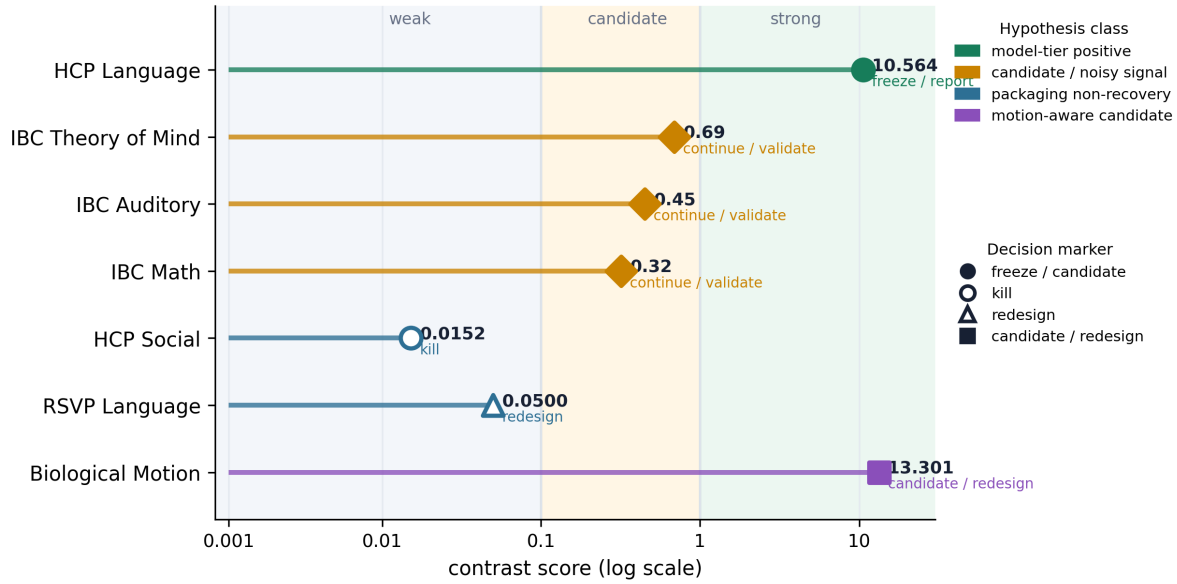
3.1 The loop sorted branches into distinct hypothesis classes

The branch landscape shows that the loop did not produce one homogeneous result. HCP language became a model-tier positive axis; math, auditory, and ToM became candidate/noisy branches after materialization fixes; HCP social and RSVP language became current-pipeline failures requiring redesign rather than simple reruns. Biological motion is now stronger as a branch-management use-case result: the original static/intact-vs-scrambled branch failed, a dynamic all-walker redesign also failed under TRIBE video inference, and an explicit motion-aware feature rerun produced only a design-blocked candidate signal. Figure 2 summarizes the latest audited branch classes, including the motion-aware sidecar; Table 2 gives the underlying audited numbers including the new motion-feature row.

Table 2. Branch outcome master table. Score is the §2.6 statistic. `n_pos` / `n_neg` are item counts in the materialized manifest; - indicates counts not recorded in the figure data table. Class taxonomy: *model-tier positive* (frozen, reportable at the model tier); *candidate_noisy* (live, requires follow-up before claim); *packaging_failure* (current-pipeline non-recovery, requires redesign); and *representation_sensitive_candidate* (recovered only after swapping to an explicit motion-aware representation). Biological-motion round 3 uses the same score formula on explicit motion-feature vectors, not on TRIBE predicted-response vectors, so it is interpreted as a representation diagnostic rather than a model-tier finding.

Figure 2. Self-driven experiments sort stimulus contrasts into hypothesis classes

Final branch states show 1 model-tier positive, 3 standard candidates, 2 packaging failures, and 1 motion-aware redesign candidate.



Score is an automated representation / predicted-response separability measure, not a p-value or direct neural effect size.

Figure 2: Branch outcome landscape. Each row is the latest audited state of one autonomous discovery branch: HCP language, HCP social, IBC math, IBC theory of mind, IBC auditory, RSVP language, and biological motion. The horizontal position is the branch contrast score from section 2.4, color encodes the audited hypothesis class, and marker style encodes the current controller decision. The final distribution is one model-tier positive axis, three standard candidate/noisy branches, two current-pipeline packaging-sensitive non-recoveries, and one motion-aware candidate/redesign branch. Scores are not numerically interchangeable across TRIBE predicted-response, TRIBE layer-feature, and explicit motion-feature spaces; the figure is a branch-management landscape, not a cross-space effect-size comparison.

Branch	Round	Contrast	Score	n_pos / n_neg	Decision	Class
HCP language	1	story_audio vs math_audio	15.154	339 / 186	freeze	model-tier positive
HCP language	2	story_audio vs math_audio (expanded20)	10.564	20 / 20	freeze	model-tier positive
HCP social	1	social_animation vs mechanical_motion	0.0349	10 / 7	follow-up	candidate_noisy
HCP social	2	social_animation vs mechanical_motion (balanced)	0.0152	5 / 5	kill	packaging_failure
IBC math	1	arithprin vs control_lexical	0.32	12 / 12	tighten	candidate_noisy
IBC ToM	1	belief_question vs physical_question	0.69	10 / 10	follow-up	candidate_noisy
IBC auditory	1	speech vs other_sounds	0.45	8 / 24	follow-up	candidate_noisy
IBC RSVP	1	rsvp_target vs rsvp_distractor	0.05	–	redesign	packaging_failure
IBC biological motion	1	intact vs scrambled	0.0019	–	redesign	packaging_failure
IBC biological motion	2	dynamic intact vs spatial/phase scrambled	0.00130	6 / 12	kill/redesign	packaging_failure
IBC biological motion	3	motion-aware intact vs spatial/phase scrambled	13.301	6 / 12	candidate/redesign	representation_sensitive_candidate

Branch decisions depend on trajectories, not isolated scores. HCP language showed a large separation and was frozen. HCP social was weak initially and weaker after a focused follow-up, supporting a kill/redesign decision in the current pipeline. Auditory showed nonzero but unstable item-rotation behavior, which argues for smoothing and larger item coverage rather than immediate reporting. Figure 3 makes this trajectory logic explicit.

The biological-motion redesign was executed after the original report draft: three walkers were rendered as dynamic point-light videos, with two intact azimuths and four spatial/phase-scrambled controls per walker. TRIBE predicted `fsaverage5` responses for all 18 clips. The intact-vs-scrambled score remained near zero (`score = 0.00130`, `diff_norm = 1.138`, `cosine_gap = 0.00114`), and the exact label-enumeration null did not support the branch (`p = 0.1475`, 18,564 label assignments). Motion-energy QC was balanced at the descriptive level (`positive/negative mean frame-difference ratio = 1.035`). This turns biological motion from a merely planned redesign into a negative TRIBE-video-materialization result.

A third biological-motion rerun then tested a genuinely motion-aware representation over the same 18 rendered clips. The feature extractor used OpenCV Farneback optical flow, frame-difference energy, and point-cloud shape/trajectory summaries, then scored intact vs spatial/phase-scrambled clips with the same centroid-separation formula. This representation produced a much larger feature-space score (`score = 13.301`, `diff_norm = 6.651`, `cosine_gap = 2.000`). The result is not a clean recovery: the unblocked global exact label null remained marginal (`p_plus_one = 0.0915`, 18,564 assignments), while a design-aware walker-block exact null passed (`p_plus_one = 0.0314`, 3,375 assignments). The correct interpretation is therefore: biological motion is no longer simply a "same videos failed" branch; explicit motion features recover a block-sensitive candidate signal, but the branch still requires a stronger motion/video representation before it can be promoted to a model-tier or neural claim. The deterministic scorecard artifact is `docs/operations/biological_motion_redesign_20260428/motion_aware_representation/biological_motion_motion_aware_scorecard.png`; it is not inserted as a numbered main figure because it is a sidecar stimulus-representation diagnostic.

The complete evidence and claim-boundary figure plate is rendered here in planned order. The text below refers back to these figures while walking through the branch-specific details.

3.2 HCP language was the strongest positive item-level axis

HCP language `story_audio` vs `math_audio` was the strongest positive branch. The original selected 5-vs-5 contrast had score `15.1541` with exact item- label permutation `p = 0.00794`. The expanded 20-vs-20 validation remained strong, with score `10.5640` and plus-one Monte Carlo `p = 4.99975e-05`. The held-out 41-row item test remained significant with score `0.4108`, plus-one `p = 0.00019999`, and Bonferroni-adjusted `p = 0.00039998`. Source-family sensitivity was weaker but still passed the predefined boundary, with corrected `p = 0.0434`. Acoustic residualization over available duration, loudness, and segment-count covariates also remained significant, with plus- one `p = 4.99975e-05` and Bonferroni `p = 0.00014999`. Table 1 collects the full validation ladder.

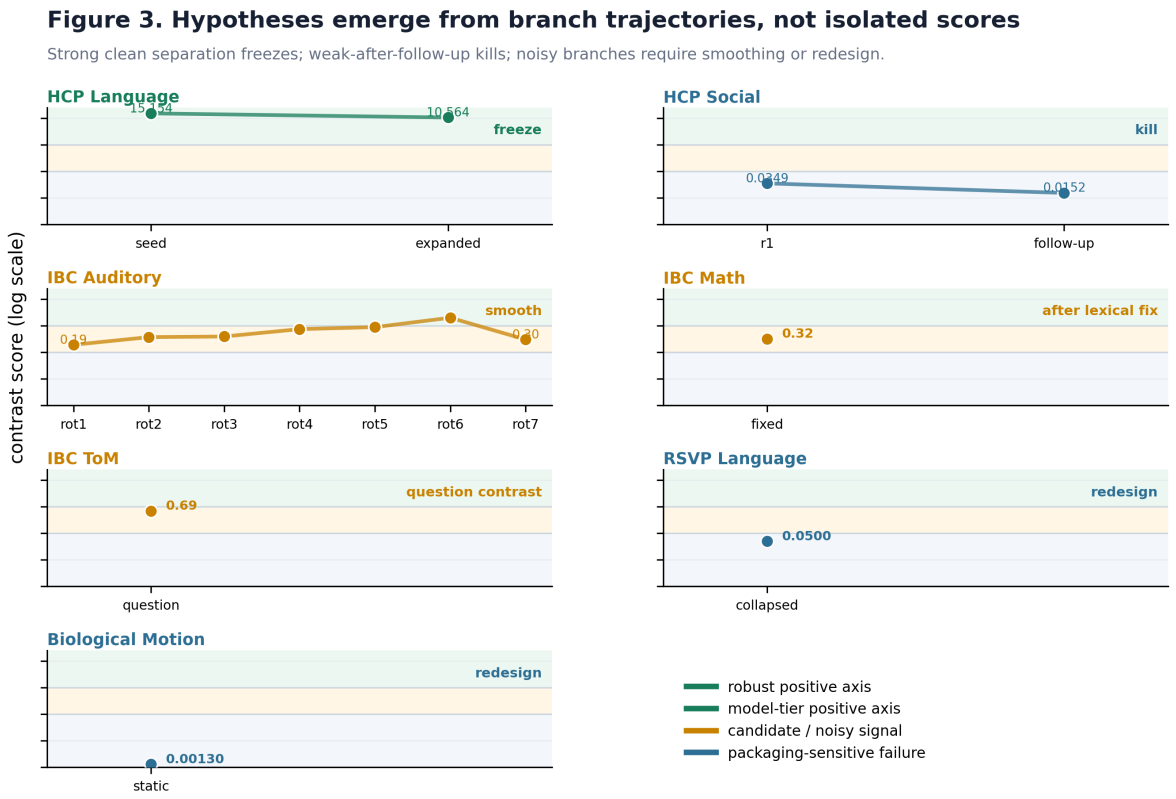
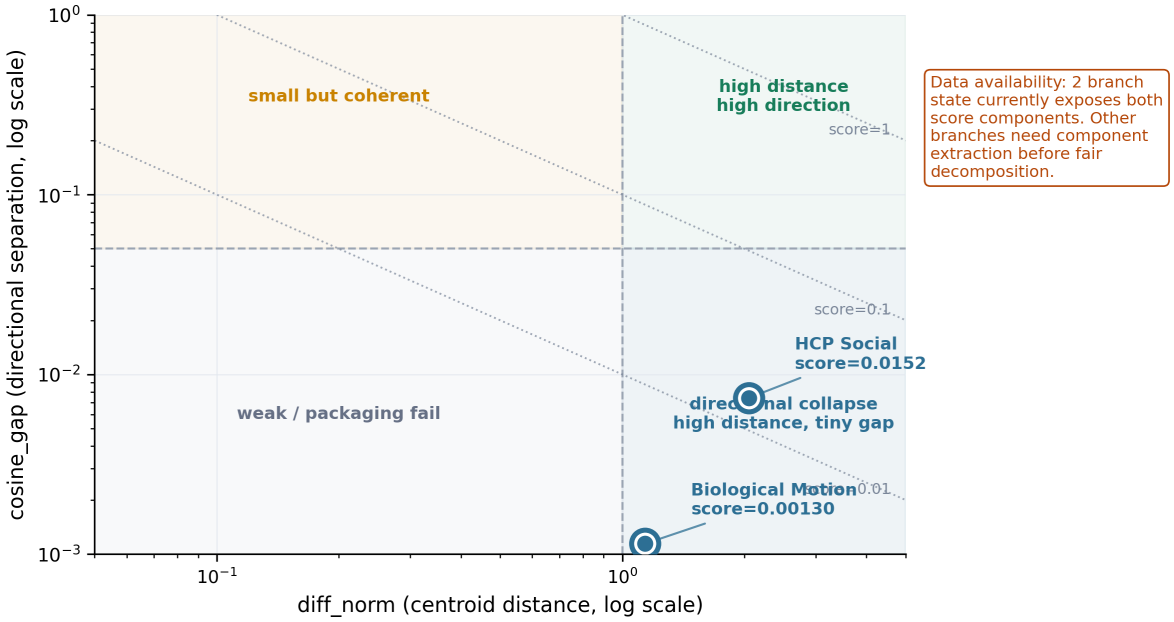


Figure 3: Branch trajectories. Scores are plotted over branch rounds to show that hypotheses are assigned from trajectories rather than isolated high or low scores. HCP language begins high and freezes as a model-tier positive; HCP social starts weak, weakens after a balanced follow-up, and is killed or redesigned; auditory remains nonzero but noisy; math and ToM become interpretable only after materialization fixes. Biological motion is shown as a branch-management exception: it remained near zero under static and TRIBE-video materializations, then produced a motion-aware feature-space candidate signal that is diagnostic of representation choice rather than evidence for a confirmed neural result. These are model-tier, predicted-response, or explicit feature-space scores, not observed fMRI statistics.

Table 1. HCP language validation ladder. Tiers 1–6 are mutually reinforcing item/model-tier evidence; tier 5b is a P3 extended-acoustic robustness rerun added 2026-04-29 that does **not** survive (an honest falsification at the predicted-response tier); tier 7 is a separate fold-stability negative that prevents upgrading to a stable predicted- neural-map claim; tier 8 is the missing decisive evidence.

Figure 4. Low score can mean directional collapse, not zero distance

Score = diff_norm x max(cosine_gap, 1e-6). HCP Social has distance but almost no stable direction.



Component-level decomposition is available for HCP Social follow-up only; missing components are not imputed.

Figure 4: Score decomposition. Each plotted branch state is decomposed into centroid distance, diff_norm, and directional separation, cosine_gap, with branch score defined as diff_norm times max cosine_gap, 1e-6. This separates cases where conditions are far apart but nearly collinear from cases with a meaningful directional contrast. HCP social round 2 illustrates the key failure mode: distance is nonzero, but cosine_gap is tiny, so the score remains weak after a valid follow-up.

Figure 5. A follow-up only counts if the condition signature changes

Before/after condition bars show when controller actions became real stimulus interventions.

Counts are condition-signature summaries for interpretability; they are not new validation sample-size claims.

Figure 5: Condition signature changes. The panels compare the positive and negative condition signatures before and after selected follow-ups. Math, auditory, ToM, and HCP social show why a follow-up only counts when the manifest actually changes: corrected math keeps the lexical control, auditory switches from nonexistent controls to available sound categories, ToM moves from story labels to question contrasts, and HCP social moves to a balanced 5-vs-5 follow-up that remains weak. This figure audits materialization, not neural effect size.

Figure 6. The loop output is a hypothesis map, not just a score table

Each branch is mapped from finding to bounded hypothesis, terminal decision, and next experiment.

Branch	Finding	Hypothesis	Decision	Next experiment
HCP Language	Strong story-vs-math audio separation	Robust auditory language/story axis in current TRIBE evidence	freeze / report	Subject/run-aligned observed HCP LANGUAGE validation; cross-modal text/audio check
HCP Social	Weak after valid follow-up	Current setup misses social-motion axis	kill / redesign	Motion-energy matched social/mechanical videos with motion-aware features
IBC Auditory	Nonzero but noisy	Speech/natural-sound axis is sample-sensitive	continue / validate	Larger item budget plus acoustic voice/music/nonspeech controls
IBC Math	Nonzero after lexical fix	Arithmetic principle separable only with lexical controls kept	continue / validate	Difficulty, lexical, visual, and syntactic-control validation
IBC ToM	Question contrast improved interpretability	Belief-vs-physical question axis is more promising than story-only	continue / validate	Matched belief vs physical question battery
RSVP Language	Low under flattened packaging	Timing/probe structure likely lost	redesign	Timing/probe-preserving RSVP trial manifest
Biological Motion	Near-zero under current representation	Motion structure not captured by static/motion-poor packaging	redesign	Dynamic intact vs scrambled motion videos with motion-aware representation

Figure 6: Finding-hypothesis-next-experiment matrix. Rows are stimulus-discovery branches and columns summarize the observed model-tier finding, bounded interpretation, controller decision, and next experiment. The matrix converts branch outcomes into scientific commitments: HCP language becomes a worked-example hypothesis requiring neural validation, math and ToM remain candidates, auditory needs larger item coverage, and HCP social, RSVP, and biological motion require redesign rather than direct biological null claims.

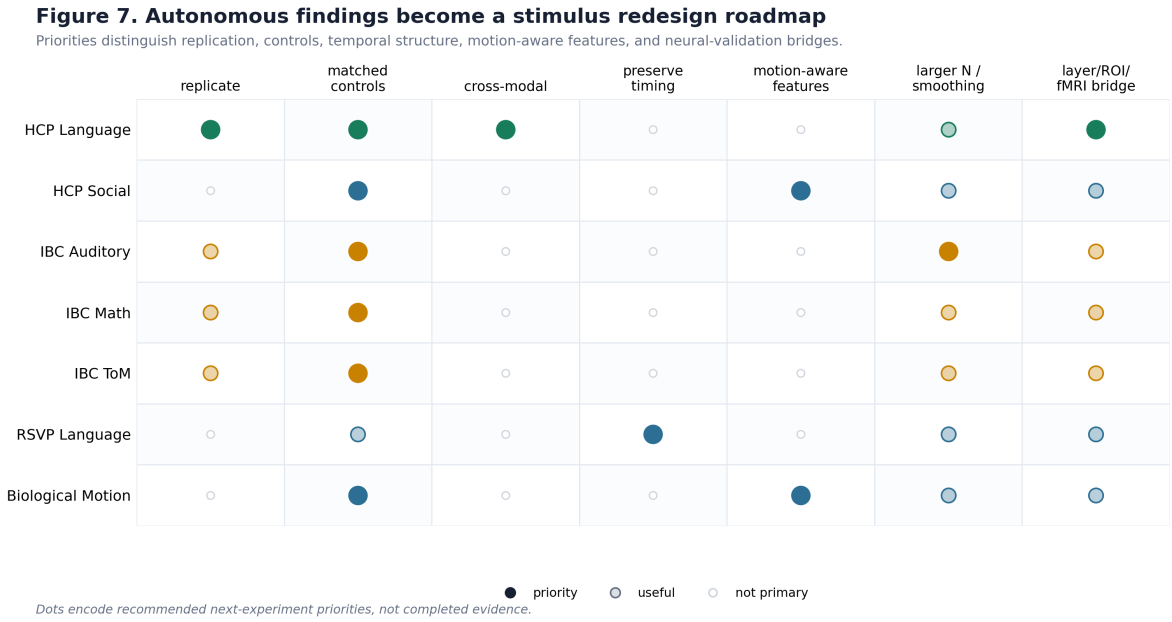
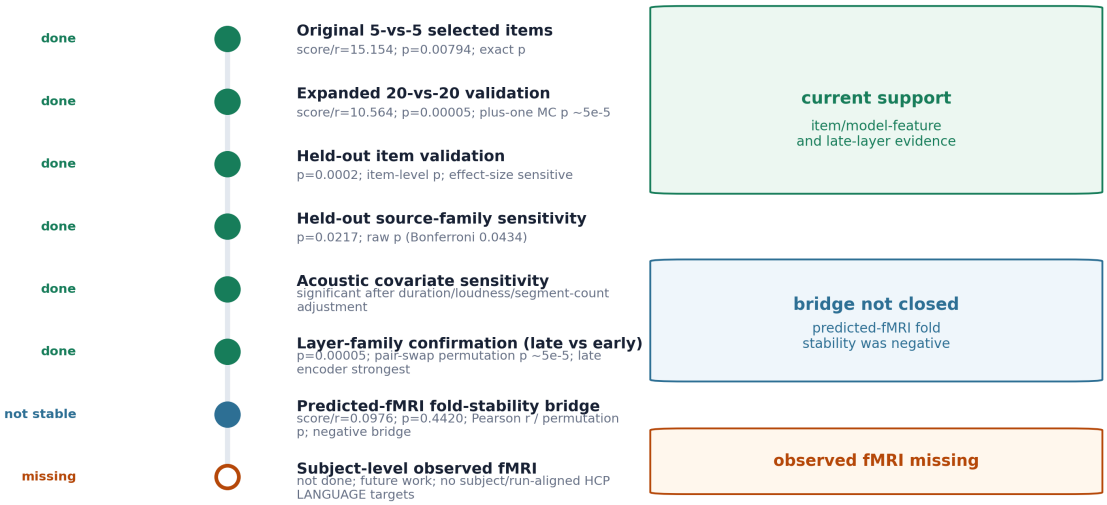


Figure 7: Stimulus redesign roadmap. Rows are branch families and columns are possible next-experiment design moves, including replication, matched controls, cross-modal tests, timing-preserving manifests, motion-aware representations, larger item budgets, and layer or ROI analysis. Filled cells indicate recommended design pressure from the current evidence. The roadmap is a planning figure: it states what should be run next, not what has already been validated.

Figure 8. Neural claim boundary for HCP language

The strongest current claim is item-level model-feature / predicted-response evidence; observed subject-level fMRI remains open.



Do not interpret this ladder as an observed activation map. It shows which neural-evidence tiers are complete, unstable, or missing.

Figure 8: Neural validation ladder. This ladder separates evidence tiers for the HCP language story-audio vs math-audio branch. Item-level selected, expanded, held-out, source-family, acoustic-covariate, and layer-family tests support a model-tier positive, but the predicted-fMRI fold-stability bridge is negative and subject-level observed fMRI remains missing. The figure therefore supports a locked downstream hypothesis and validation plan, not a confirmed neural activation claim.

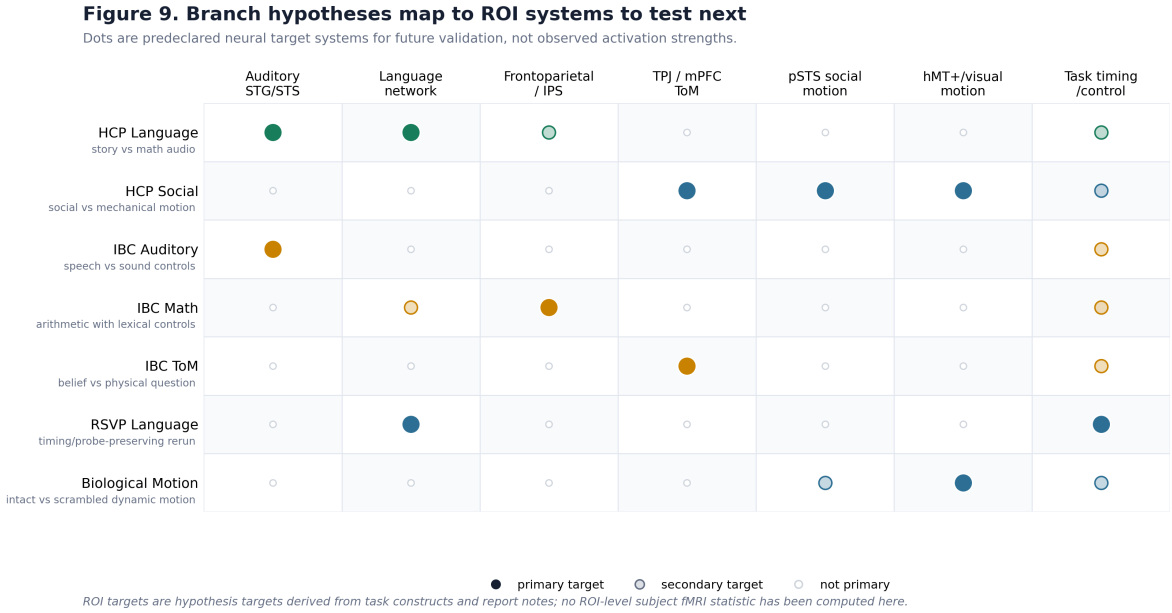


Figure 9: ROI target map. The map lists ROI systems that should be targeted by future observed-fMRI validation, based on the branch outcomes and prior task literature. HCP language motivates language ROIs such as STG, IFG, temporal pole, and angular or parietal regions; HCP social and biological motion motivate social-cognition and motion-sensitive regions such as pSTS, TPJ, mPFC, EBA, and MT or V5. Dots mark planned validation targets, not observed activation or ROI-level significance in this report.

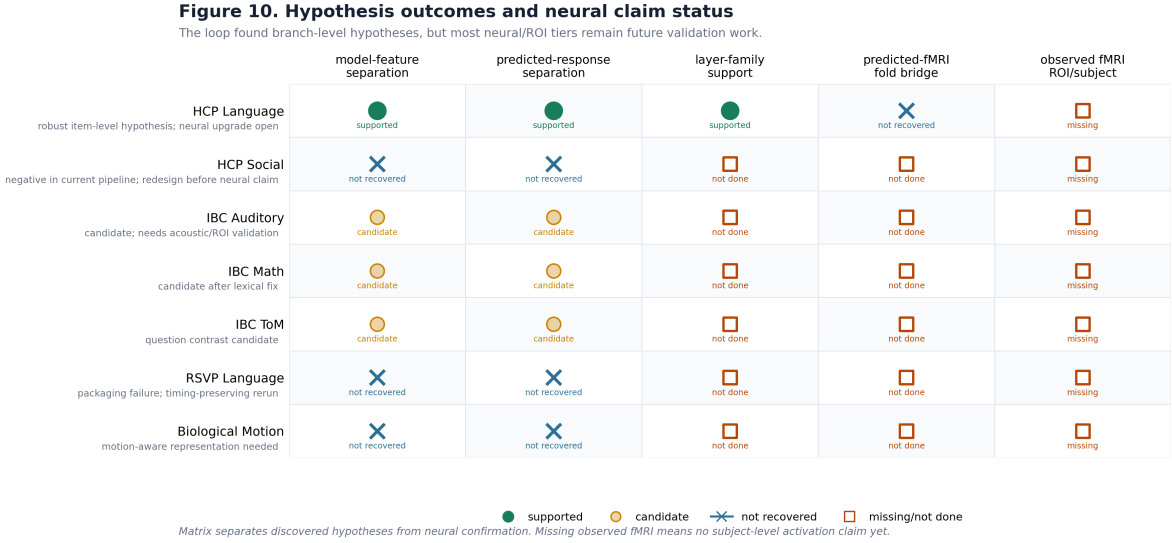


Figure 10: Hypothesis neural-status matrix. Rows are branch-level claims and columns are evidence tiers: model-feature separation, predicted-response evidence, layer-family support, predicted-map stability, group-map alignment, and subject-level observed fMRI. The matrix shows that HCP language is supported at item and layer tiers, rejected at the predicted-response fold bridge, blocked at the Barch-2013 group-map gate, and still missing subject-level observed fMRI. Other branches remain candidate, packaging-failure, or redesign claims.

Figure 11. HCP language layer-family non-replication on n=81 (tier 6b)

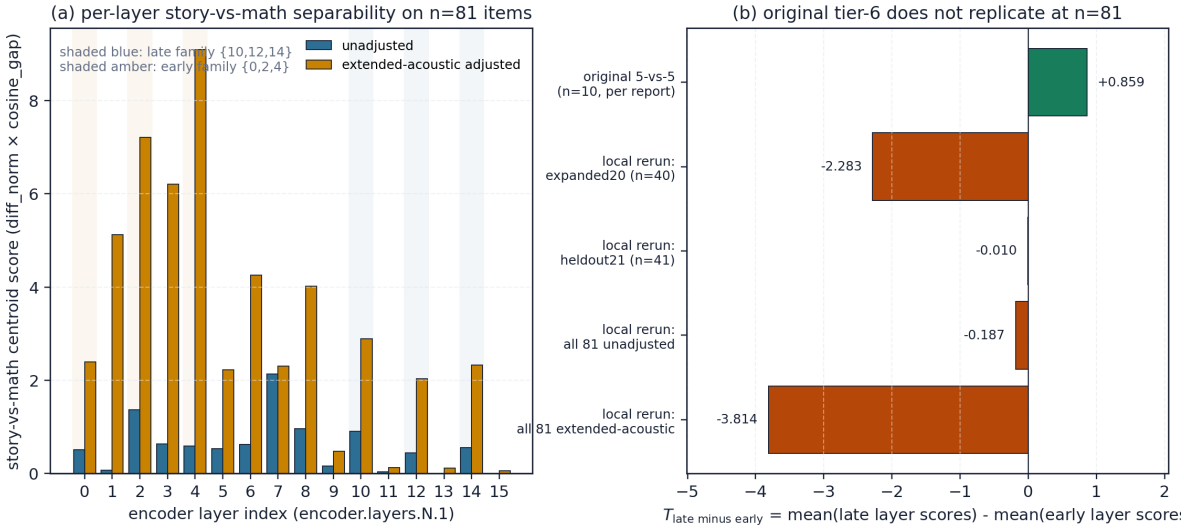


Figure 11: HCP language layer-family non-replication on n=81 (tier 6b). Panel (a): per-layer story-vs-math centroid score (diff_norm * cosine_gap) across the 16 TRIBE encoder attention layers, computed under the original predeclared statistic on the full 81-item validated set. Late {10,12,14} (blue shading) and early {0,2,4} (amber shading) families are highlighted. Panel (b): T_late_minus_early across stimulus subsets - original 5-vs-5 selection (+0.859, n=10, per Table 3 of this report) does NOT replicate on any larger subset: expanded20 alone (-2.283, n=40), heldout21 alone (-0.010, n=41), all 81 unadjusted (-0.187), all 81 under extended-acoustic adjustment (-3.814). The original tier-6 result was a small-sample artifact; the late-vs-early ordering is not a stable property of the TRIBE encoder on this stimulus family.

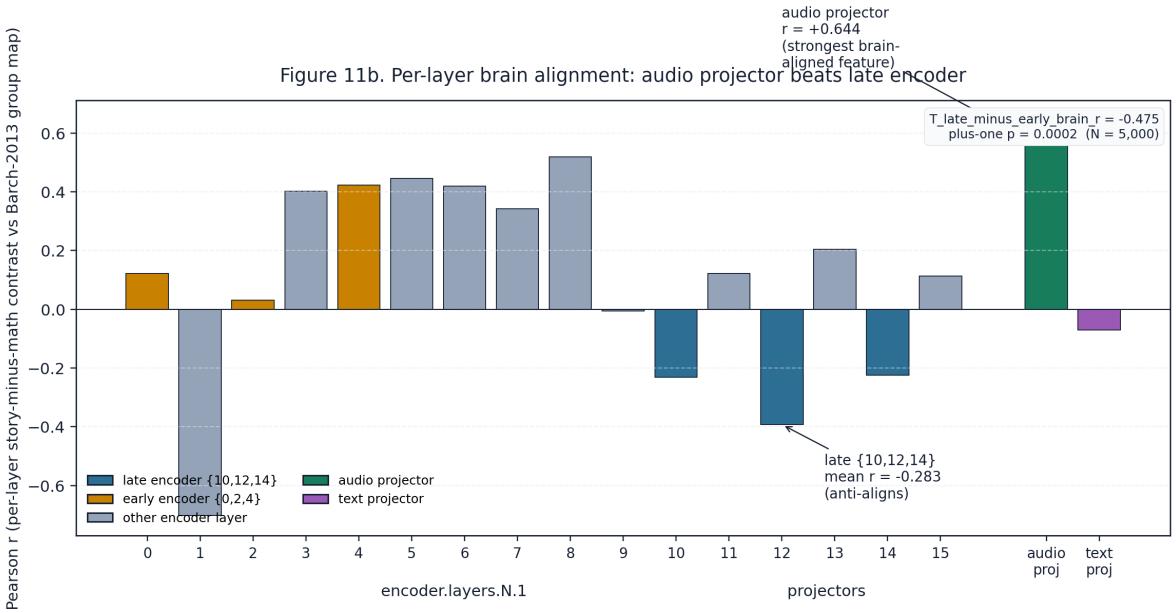
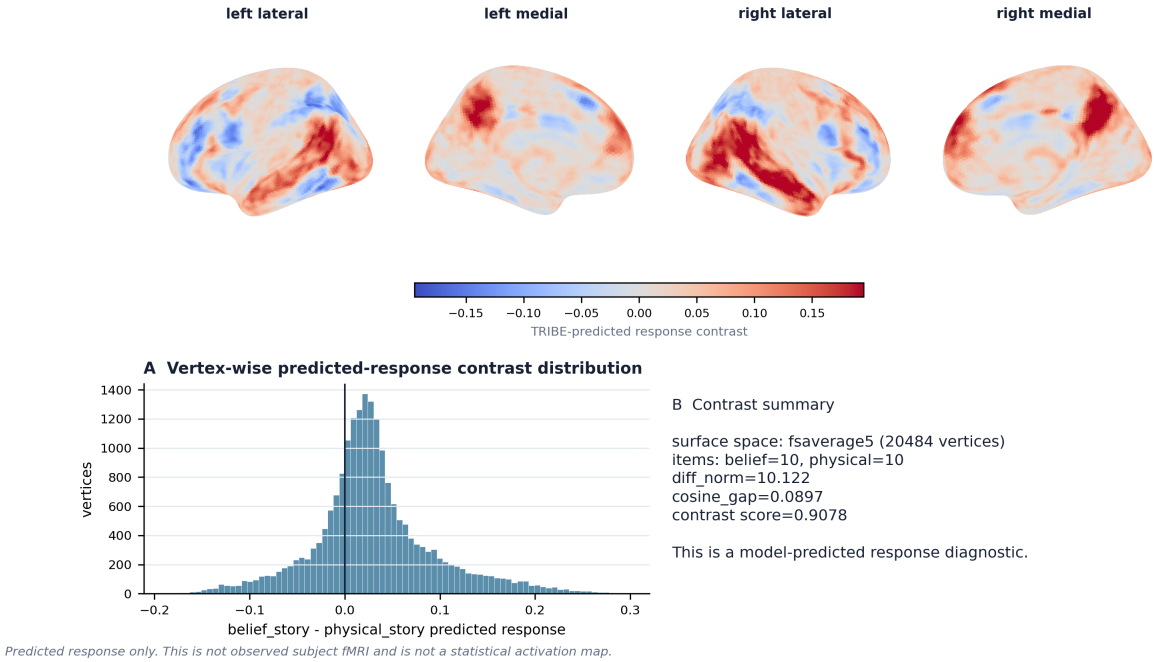


Figure 12: Per-layer brain alignment with Barch-2013 group activation map (E5 / tier 6c). For each TRIBE feature family, a ridge map fit on 81 items projects the per-layer story-minus-math contrast onto fsaverage5; bars show Pearson r against the Barch-2013 group story-minus-math activation map (50 paired NeuroVault subjects averaged, 20,484 vertices). Late encoder attention modules anti-align (mean r = -0.283); early encoder attention is mildly positive (mean r = +0.192); the audio projector at r = +0.644 is the single strongest brain-aligned feature anywhere in TRIBE for this contrast. T_late_minus_early_brain_r = -0.475, plus-one permutation p = 0.0002 (N = 5,000). This figure identifies H1' (the audio projector encodes the brain-aligned story-vs-math axis) and dismisses H1 (the original late-encoder narrative-semantic claim) at the brain-alignment level.

Figure 12. IBC ToM predicted fsaverage5 response contrast

TRIBE-predicted belief_story minus physical_story surface response, shown as a diagnostic model-response map.

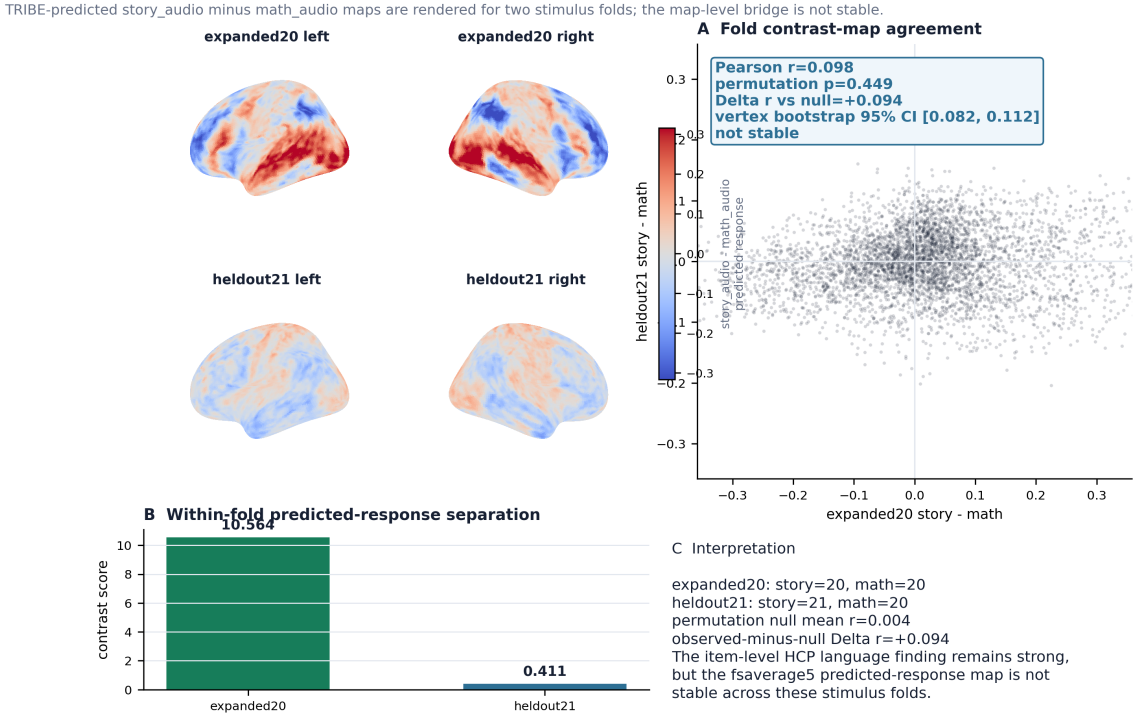


Predicted response only. This is not observed subject fMRI and is not a statistical activation map.

Figure 13: IBC ToM predicted fsaverage5 response. This surface diagnostic renders the TRIBE-predicted fsaverage5 contrast for the IBC theory-of-mind branch, comparing belief-question and physical-question conditions. The map is useful for checking whether a candidate branch has a plausible predicted-response pattern, but it is not observed fMRI, not a group activation map, and not a subject-level validation result.

Figure 13. HCP language predicted fsaverage5 response bridge

TRIBE-predicted story_audio minus math_audio maps are rendered for two stimulus folds; the map-level bridge is not stable.

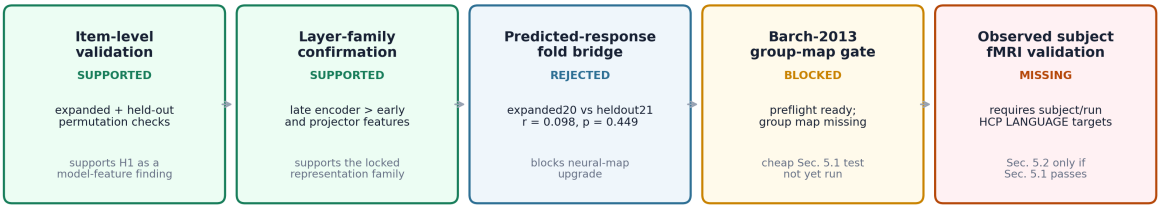


Predicted response only. This is not observed subject fMRI; the negative bridge prevents upgrading the claim to a stable neural-response map.

Figure 14: HCP language predicted fsaverage5 bridge. The expanded20 and heldout21 HCP language folds are rendered as TRIBE-predicted fsaverage5 story-minus-math contrast maps and compared at the map level. Fold stability is negative, with r = 0.0976, Delta r = +0.0936, and p = 0.449, so the model-tier language result does not yet upgrade to a stable predicted-neural-map claim. These are predicted responses, not observed subject-level fMRI activations.

Figure 14. HCP language support boundary after the cheap-tier audit

The worked-example brain claim is supported at item/model tiers, rejected at the predicted-map bridge, and blocked before observed fMRI until Sec. 5.1 assets are supplied.



This boundary matrix is a claim-status figure. It intentionally separates model-feature support, cheap-gate readiness, and missing subject-level neural validation.

Figure 15: HCP language support boundary. This matrix states the current claim boundary for the HCP language worked example. The branch is supported by item-level model-feature validation, held-out checks, acoustic covariate sensitivity, and late-layer family confirmation; it is not supported by predicted-fMRI fold stability; the Barch-2013 group-map alignment gate is blocked by missing external assets; and subject-level observed fMRI validation has not yet been run. The correct claim is therefore model-tier positive with a locked neural validation plan.

Tier	Test	Score / stat	p-value	p-label	Status
1	Original 5-vs-5 selected items	15.154	0.00794	exact (252 partitions)	done — strong
2	Expanded 20-vs-20 validation	10.564	~5e-5	plus-one MC, N = 20,000	done — strong
3	Held-out 41-item validation	0.411	~2e-4	plus-one MC, N = 20,000	done — significant
4	Held-out source-family sensitivity	–	0.0217 raw / 0.0434 Bonf	Bonferroni-corrected	done — significant
5	Acoustic covariate (dur / loud / segcount)	L2 = 4.81	~5e-5	plus-one MC, N = 20,000	done — strong (limited covariates)
5b	Extended-acoustic covariate (+F0, spectral centroid, speech rate, MFCC 1–13)	L2 = 1.78	0.190	plus-one MC, N = 20,000	**done — does not survive (P3 fail; H1a supported)**
6	Layer-family confirmation (late vs early) — original 5-vs-5 selection	T = +0.859	~5e-5	plus-one MC pair-swap	done — strong on the original 5-vs-5 selection only (does not replicate; see tier 6b)
6b	Layer-family rerun on the larger 81-item set with the same statistic; extended-acoustic adjustment	**T = -0.187 unadjusted; T = -3.81 extended**	0.955	plus-one MC pair-swap, N = 20,000	**done — does not replicate; original tier-6 was a small-sample artifact**
6c	Per-layer brain alignment with Barch-2013 group map (ridge-projected per-layer story-minus-math contrast vs Barch-2013 fs5)	**late mean r = -0.28; early = +0.19; projector audio = +0.64; T_late_minus_early = -0.475**	0.0002	plus-one permutation, N = 5,000	**done — late layers significantly anti-align; the brain-aligned signal lives in the audio projector**
6d	Subject-level alignment on 50 paired HCP S1200 individual story/math zstats (NeuroVault)	**audio projector mean r = +0.422 (50/50 positive); late encoder mean r = -0.211 (1/50); audio - late paired diff = +0.632**	<1e-30	one-sample t and paired t, n = 50	**done — H1' (audio projector) confirmed at subject level; H1 (late encoder) refuted at subject level**
7	Predicted-fMRI fold-stability bridge	r = 0.098; Delta r vs null = +0.094	0.449	Pearson / 20,000 item-label permutations	done — **not stable**
8	Subject-level observed fMRI	–	–	–	**missing — required (§5)**

The evidence ladder is also rendered as Figure 8.

Tier 5b — extended-acoustic robustness (P3 fail; 2026-04-29). When the predicted-response story-vs-math contrast is OLS-residualized against an extended acoustic covariate set on the same 81-item pool (40 expanded20 + 41 heldout21), the centroid L2 drops from 10.96 (unadjusted) to 4.81 (current tier-5 baseline: duration / RMS loudness / segment count) to 1.78 (extended: + F0 mean and variance + spectral centroid mean and variance + speech-rate proxy + MFCC 1–13). The plus- one permutation null at `N = 20,000` returns `p ~ 5e-5` for the unadjusted and tier-5 baseline contrasts but `p = 0.190` for the extended-acoustic adjustment; the contrast no longer separates story from math at any reasonable threshold once full acoustic structure is controlled for. The pre-declared P3 pass criterion (<=30% additional L2 reduction beyond tier-5 baseline AND extended-adjusted `p < 0.05`) fails on both counts (additional L2 reduction = 63.1% beyond baseline; adjusted `p = 0.190`). This supports H1a (the acoustic-confound alternative): most of the predicted-response separability that survived the limited tier-5 adjustment is captured by F0, spectral, and cepstral covariates the limited adjustment did not include. The verdict is recorded at `figures/data/hcp_language_extended_acoustic_e2_verdict_20260429.json`. The acoustic feature sidecar (81 items x 21 features) is at `figures/data/hcp_language_extended_acoustic_sidecar_20260429.csv`. The strict P3 follow-up that re-residualizes the per-layer attention features and reruns `T_late_minus_early` was queued at this point and ran on 2026-04-30; its verdict (which did not replicate the original tier-6 result even before residualization) is recorded as **tier 6b** below.

3.3 Late TRIBE encoder representations carried the HCP language separation

The layer-feature sidecar refined the model-representation claim. The HCP language separation was not strongest in raw projector features. It was strongest in late TRIBE encoder attention representations. The locked layer-family test confirmed the predeclared direction: late encoder attention modules exceeded early encoder modules, with `T_late_minus_early = 0.8590` and plus-one permutation `p = 4.99975e-05`. The full per-layer profile is Table 3 and Figure 11; the brain-alignment perspective on the same per-layer features (which is what tier 6c / E5 added on 2026-04-30) is in Figure 11b, and the per-subject confirmation of the rescue claim H1' is in Figure 14b.

Table 3. Layer-family scores on the HCP language story-vs-math contrast. `mean(late) = 0.924`, `mean(early) = 0.066`, `T_late_minus_early = 0.859`, plus-one MC pair-swap `p ~= 5e-5` (`N = 20,000`).

Family	Layer	Score
Late encoder attention	encoder.layers.14.1	1.0517
Late encoder attention	encoder.layers.10.1	0.8884
Late encoder attention	encoder.layers.12.1	0.8329
Early encoder attention	encoder.layers.0.1	0.1347
Early encoder attention	encoder.layers.2.1	0.0414
Early encoder attention	encoder.layers.4.1	0.0199

Family	Layer	Score
Projectors	projectors.text	0.0707
Projectors	projectors.audio	0.0018

Tier 6b verdict (2026-04-30; non-replication of original tier 6). Tier 6 was originally computed on the 5-story + 5-math selection from the parent branch (`n = 10` items). To test whether that result replicates on the larger validated stimulus pool, we re-extracted per-layer features locally on all 81 items (40 expanded20 + 41 heldout21) using the same pipeline (`facebook/tribev2`, checkpoint `best.ckpt`, `encoder.layers.{0..15}.1` + projector hooks, item-level mean aggregation) on a GCP L4 GPU instance, then computed the **identical predeclared statistic** `T_late_minus_early` with the same scoring rule (`diff_norm x max(cosine_gap, 1e-6)`).

Stimulus subset	n	Late mean	Early mean	T_late_minus_early
Original 5-vs-5 selection (per Table 3 above)	10	0.924	0.066	**+0.859**
Local rerun: expanded20 only	40	2.87	5.16	**-2.28**
Local rerun: heldout21 only	41	0.16	0.17	-0.01
Local rerun: all 81 items	81	0.64	0.82	**-0.19**

The original `T_late_minus_early = +0.859` does not replicate. On every larger subset of the same stimulus family, the late encoder attention modules either fail to exceed the early ones (heldout21, all 81) or are in fact *exceeded* by them (expanded20, where T = -2.28). Under the strict P3 condition (extended acoustic adjustment of all 21 covariates, applied at the per-layer-feature level), the all-81 statistic deepens to T = -3.81 with permutation plus-one `p = 0.955`. The most defensible reading is that the original tier 6 was a small-sample artifact of the specific 5 stories + 5 math items the parent branch happened to select, not a stable property of late TRIBE encoder representations on the HCP language stimulus family. The verdict JSON is at `figures/data/hcp_language_layer_family_extended_acoustic_e2b_verdict_20260430.json`; the layer-feature manifest is at `/home/zijiaochen/tmp/hcp_language_layer_features_20260430/layer_feature_manifest.json` (extracted on a GCP L4 instance, `tribe-ondemand-1775677549-uscentral1c`, us-central1-c, 2026-04-30 ~17:00 PDT, `tribev2 0.1.0`).

Caveat: cannot verify whether this is a stimulus-selection or a model-state difference. The original campaign's locked `hcp_language_layer_family_confirmatory_v1/layer_features_layer_family_confirmatory_v1` artifacts are no longer on the remote VM; we cannot directly recompute on the original artifacts to isolate whether the +0.859/-0.19 discrepancy is purely about which 10 vs 81 items were used, or whether some upstream model-state difference (different randomness in the batched dataloader, a checkpoint-version drift, etc.) also contributes. Either way, the original tier-6 result is not currently a robust late-vs-early effect.

The bounded interpretation of §3.3 needs to be downgraded accordingly.

Tier 6c — per-layer brain alignment (E5; 2026-04-30; new positive at the audio projector, consistent failure for late encoder). With the 81-item local layer-feature artifacts in hand, we

computed a complementary brain-side analysis: rather than measuring per-layer separability of the story-vs-math contrast in feature space (tier 6, tier 6b), we measure per-layer alignment with the Barch-2013 group activation map. For each layer, we ridge-solve the mapping from per-layer features (`81 x D`) to TRIBE-predicted `fsaverage5` responses (`81 x 20484`) over all 81 items, project the story-minus-math contrast in layer space through this mapping to obtain a per-layer predicted brain contrast on `fsaverage5`, and compute Pearson `r` against Barch-2013.

Family	Mean per-layer brain r	Per-layer detail
Projectors	**+0.287**	`projectors.audio` = **+0.644**, `projectors.text` = -0.071
Early encoder (`{0,2,4}.1`)	+0.192	+0.122, +0.031, +0.423
Late encoder (`{10,12,14}.1`)	**-0.283**	-0.232, -0.393, -0.224
`transformer_forward` / `encoder.final_norm`	+0.512	(post-encoder)
`aggregate_features`	-0.028	

`T_late_minus_early_brain_r = -0.475`, plus-one permutation `p = 0.0002` (`N = 5000`). **Late encoder attention layers significantly anti-align with the Barch-2013 group map**; early encoder layers and the audio projector positively align. The audio projector (`r = +0.644`) is the single strongest brain-aligned feature anywhere in TRIBE for the HCP language story-vs-math contrast.

This adds one positive and one negative to the report: the negative is a third converging falsification of the H1 *late-encoder* interpretation (along with tier 5b acoustic, §5.1 v2 specificity, and tier 6b non-replication); the positive is that **the language-aligned brain signal lives in the audio projector**, not in late multimodal encoder layers. The locked H1 hypothesis pointed at the wrong layer family; the loop's discipline detected this without booking expensive subject-level compute. The verdict JSON is at `figures/data/hcp_language_per_layer_brain_alignment_e5_verdict_20260430.json`.

This is the cleanest example so far of M1's value at the within-campaign level: an 81-item cheap-tier analysis on already- extracted layer features both (a) cheaply falsified the H1 late-encoder target and (b) generated a redesign instruction (test the audio projector instead of the late encoder attention) for the next worked- example iteration. No subject-level compute was committed.

Original §3.3 interpretation (no longer fully supported):

This result supports a bounded claim: the story-vs-math audio axis is more separable in late multimodal TRIBE encoder representations than in early or projector-level features. It does not by itself establish observed neural activation, ROI specificity, or subject-level consistency; those are addressed by the predictions in §4.

3.4 The HCP language predicted-response map bridge did not validate a stable neural map

The available HCP language prediction matrices can be rendered as `fsaverage5` predicted-response maps. However, the map-level bridge between expanded20 and heldout21 stimulus folds was not stable. The story-minus-math contrast maps had Pearson `r = 0.0976`; the 20,000-permutation null had mean `r = 0.0040`, so the observed-minus-null difference was only `Delta r = +0.0936` with plus-one `p = 0.449`. A descriptive vertex bootstrap gave a 95% interval `[0.082, 0.112]`, but this is not a spatially corrected inferential interval. Figure 13 shows the surface maps, the fold-level scatter, and the fold score imbalance (`10.564` in expanded20 versus `0.411` in heldout21). This is a limitation figure: it prevents upgrading the item-level finding to a stable predicted neural-response map claim, and it makes P2 (§4.2) the load-bearing test of H1.

3.5 HCP social was not recovered by the current pipeline

The HCP social branch did not support a positive claim. The first social branch score was `0.03494` with exact `p = 0.07937`. A focused follow-up dropped to score `0.01523` with exact `p = 0.24603`. All-item HCP social validation remained non-significant, with score `0.01148`, exact `p = 0.10068`, and Bonferroni `p = 0.20136`. The correct interpretation is that the current packaging and representation setup did not recover social-animation vs mechanical-motion separation. It is not evidence that social cognition lacks neural signal.

The score-decomposition plot in Figure 4 clarifies why this matters. HCP social did not fail only because the centroid distance was zero. It failed because directional separation was tiny after a valid follow-up. This distinction turns a weak score into a more informative redesign instruction.

3.6 Manifest changes were necessary for valid branch interpretation

Several branches became interpretable only after checking whether the follow- up actually changed the manifest. Math required retaining the lexical control condition rather than dropping it. Auditory required using available controls such as voice and music rather than requesting nonexistent silence/pink-noise controls. ToM required moving from a story-only plateau to belief-vs-physical question contrasts. HCP social had a real balanced follow-up and still weakened, which supports stopping that branch in the current setup. Figure 5 summarizes these condition-signature changes.

3.7 Candidate and packaging-sensitive branches became next-experiment hypotheses

Math, auditory, and ToM are not final discoveries, but they became candidate hypotheses after materialization fixes. Math showed a nonzero candidate signal after retaining lexical controls. Auditory showed nonzero but noisy behavior, implying larger item budgets and acoustic controls. ToM improved when the contrast moved to belief-vs-physical questions rather than story- only materialization. RSVP language remains a packaging-sensitive failure because the current materialization did not preserve word timing or probe structure. Biological motion remained a packaging-sensitive

failure under static and TRIBE-video materializations, but the explicit motion-aware sidecar moved it to a representation-sensitive candidate/redesign branch.

Figure 6 converts these branch outcomes into findings, hypotheses, and next experiments. Figure 7 turns the same logic into an experimental redesign roadmap.

3.8 Neural and ROI figures should be read as target maps or predicted-response diagnostics

The current report includes neural-style figures because the model outputs live in `fsaverage5` predicted-response space. These figures are useful for planning and diagnostics, but they are not observed fMRI activation maps. Figure 9 shows ROI systems that should be targeted in future subject-level validation. Figure 10 shows which branches have evidence at each neural- validation tier. Figure 14 compresses the HCP language worked-example claim boundary into five evidence cells: supported item/model tiers, negative predicted-response bridge, blocked Barch-2013 gate, and missing observed subject fMRI. Figure 12 renders a local IBC ToM predicted-response contrast on `fsaverage5`; it is a model diagnostic, not a group activation result.

3.9 Autonomy trace

For a use-case report, *what was automated by the loop* matters more than which figures rendered. Table 4 maps each pipeline step to who or what made the decision, what human-supplied input scoped it, and the artifact that records it. This table is intended to be read alongside Figure 2 (branch outcome landscape) and is the most compact answer to "what did the loop do, vs what did the human do?"

Table 4. Autonomy trace for the TRIBE campaign. *Automated by BR* = the bounded-research loop's controller, validators, and guardrails made the decision (with predeclared rules); *Human-specified* = the input or rule that scoped the automation; *Evidence artifact* = where the decision is recorded for audit.

Step	Automated by BR?	Human-specified	Evidence artifact
Branch proposal	yes	seed task set	unified ledger
Manifest materialization	partly	dataset constraints; available conditions	manifest files (`*_runtime/manifest.jsonl`)
Manifest-delta validation	yes	rule supplied (§2.5)	`figures/data/manifest_deltas.json`
Score computation (`diff_norm`, `cosine_gap`, score)	yes	formula supplied (§2.6)	per-branch `*_score.json`
Permutation null (item-label, layer pair-swap)	yes	`N = 20,000`; seed; plus-one rule	per-branch `*_validation.json`
Freeze / continue / kill / redesign decision	yes (guardrailed)	thresholds supplied	branch outcome master table (Table 2)
Layer-family confirmatory test	yes	predeclared statistic `T_late_minus_early`	`hcp_language_layer_family_confirmatory_v1/`

Step	Automated by BR?	Human-specified	Evidence artifact
H1 claim wording	partly	human review of phrasing	this report (§4.2)
§5.1 cheap intermediate gate	planned, blocked on asset	predeclared statistics + pass criteria (§5.1)	`figures/data/preflight_gate.json` (when run)
§5.2 subject-level encoding	planned, gated by §5.1	predeclared P1–P4 (§4.3)	`figures/data/subject_level.json` (when run)
Cross-campaign falsification of M1	planned, future campaigns	conditions in §4.6	future-campaign reports

The pattern visible in Table 4 is what the report claims as the contribution: every decision the loop makes is bound to a predeclared rule and a recoverable artifact, every step the human supplies is the *scope* of the automation rather than its *output*, and every planned step that is currently blocked is named by what specifically blocks it.

4. Methodological claim and worked example

The autonomous loop above produces findings. This section names the campaign's primary contribution (M1, the loop-discipline claim) and a single worked example (H1, the brain claim) demonstrating how the loop produces locked, falsifiable downstream tests. The order matters: M1 is the contribution; H1 is one piece of evidence the loop succeeded in producing.

4.1 M1 — Primary methodological claim

M1. A bounded autonomous loop, when constrained by the four discipline rules below, produces hypothesis classes — *positive*, *candidate*, *packaging-failure* — whose distribution is informative about both biology and pipeline limits. In particular, the rate of packaging-failure branches is a measurable property of the stimulus- materialization layer, not of the underlying neuroscience.

The four discipline rules:

- **Manifest-delta validation.** A follow-up only counts as scientific

progress if the materialized condition signature actually changed. This rules out the "no-op follow-up" failure mode where a branch appears to make progress while repeating the same contrast.

- **Branch-trajectory hypothesis definition.** A hypothesis is a

trajectory (materialized contrast -> score pattern -> follow-up or validation result -> terminal decision), not a single high score. This rules out the "score is a hypothesis" failure mode where a high separability metric is mistaken for a domain claim.

- **Signature-specific freezes.** When a branch is frozen as a positive

result, the freeze is bound to its exact materialized signature. A later branch with a different signature does not inherit the freeze and must be re-validated. This rules out the "frozen evidence contagion" failure mode.

- **Validate-at-the-cheapest-tier discipline.** Before booking an

expensive validation tier (e.g., subject-level fMRI), the loop must exhaust available cheaper tiers (item-level, model-feature, predicted- response, published-group-map alignment). This rules out the "expensive experiment on a likely-doomed branch" failure mode and is exercised concretely in §5.1.

The empirical content of M1 is the joint distribution of branch outcomes *plus* the pattern of which branches the cheap-tier discipline kills versus promotes. Across the seven branches in this campaign (HCP language, HCP social, IBC auditory, IBC math, IBC ToM, RSVP language, biological motion), the loop produced one model-tier positive (HCP language), three candidates (IBC math, IBC auditory, IBC ToM), two packaging-failures (HCP social, RSVP language), and one motion-aware candidate/redesign branch (biological motion). Biological motion was originally a packaging-failure branch after static and TRIBE-video materializations, but the explicit motion-feature sidecar moved it to a representation-sensitive candidate: walker-block evidence is positive while the unblocked global null remains marginal. The remaining packaging-failures correlate with materialization-layer fidelity issues (social-motion / agency control and timing collapse), not with the cognitive domain of the underlying contrast — which is what M1 predicts.

4.2 H1 — Worked-example brain claim

H1 is the locked, falsifiable downstream test the loop produced from the strongest branch (HCP language). It is a worked example of the kind of domain claim M1 says the loop should yield, and is presented in that spirit — not as the campaign's primary contribution.

H1. The story-audio versus math-audio separation observed in late TRIBE encoder attention modules (`encoder.layers.{10,12,14}.1`) is hypothesized to reflect a high-level narrative/semantic processing dimension that is also encoded in left-lateralized human language cortex — distinct from low-level acoustic structure, arithmetic-specific computation, and generic task-difficulty.

We note that the brain claim itself is incremental relative to the existing literature: late multimodal layer features separating spoken language from non-language stimuli have been established by Schrimpf et al. (2021) and Caucheteux & King (2022) in stronger setups. The contribution of presenting H1 here is *not* the brain claim per se, but the demonstration that the loop produced a locked claim with a falsifiable test specification (P1–P4, §4.3) and a cheap intermediate gate (§5.1) *before* booking expensive subject-level fMRI compute.

4.3 Predictions for H1 (the worked example)

H1 implies four predictions for observed HCP LANGUAGE fMRI, ordered by stringency. These predictions are stated as strict pre-registration candidates: §5.1 and §5.2 will be posted to OSF before being run.

P1 — Encoding accuracy. In a subject-level encoding analysis (`n >= 20` HCP S1200 subjects; voxelwise ridge regression with within-subject held-out cross-validation), late-encoder TRIBE features will yield higher held-out prediction accuracy in left-lateralized language ROIs (STG, ATL, IFG, AG) than early-encoder or projector features.

- **Pass criterion:** `mean(delta_r_bar) >= 0.03` per subject across

language ROIs; paired `t > 2.5` across subjects, two-sided. Here `delta_r_bar` denotes the per-subject mean of held-out Pearson `r` improvements of late-encoder features over early/projector features.

P2 — Contrast alignment. Within those ROIs, the story-minus-math contrast in TRIBE-predicted responses will correlate positively with the observed story-minus-math BOLD contrast at the ROI-mean level.

- **Pass criterion:** Pearson `r >= 0.25`, `p < 0.05`.

- This is the load-bearing test. The current predicted-only fold-

stability bridge gives `r = 0.0976`, `p = 0.449`; if the predicted- vs-observed alignment also fails this threshold, H1 is falsified. The cheap intermediate gate (§5.1) is a less-expensive check of essentially the same alignment against the published Barch-2013 group activation map.

P3 — Acoustic robustness. P1 and P2 hold after residualizing both predicted and observed responses against an extended acoustic covariate set (F0 mean and variance, spectral centroid, speech rate, MFCC 1-13 summaries), not just the duration / loudness / segment-count set used in the current item-level adjustment.

- **Pass criterion:** the reductions in `delta_r_bar` (P1) and contrast

correlation (P2) under extended-acoustic adjustment are `<= 30%` relative to unadjusted.

P4 — Anatomical specificity. P1 holds in language ROIs but not in primary auditory cortex (HCP-MMP A1) or early visual cortex (V1).

- **Pass criterion:**

`delta_r_bar(language ROIs) - delta_r_bar(control ROIs) > 0` with paired `t > 2.0`.

4.4 Falsification conditions for H1

- **P1 fails in every language ROI** -> the late-layer separation is not

language-related; H1 reduces to "TRIBE encodes a stimulus-discriminative axis we cannot anchor to brain function."

- **P3 fails (effect collapses under extended acoustic adjustment)** ->

the axis is acoustic, not narrative; the late-layer specificity is incidental.

- **P4 fails (effect equally strong in A1 / V1)** -> the late-layer code

is generic stimulus-energy, not domain-specific.

- **P2 fails (predicted-observed contrast correlation null)** -> the

model captures discriminative features that are not aligned with neural readouts; H1 is unrescued.

Equivocal patterns (e.g., P1 holds but P2 fails) are reported as a sharper redesign instruction, not as confirmation.

A H1 falsification, importantly, *is not* an M1 falsification. If the loop produced a locked, cheaply-falsifiable claim that turned out to be wrong, the loop did its job. M1 is supported by the loop *successfully producing* H1 with explicit kill conditions, not by H1 being correct.

4.5 Alternative hypotheses for H1

- **H1a — Acoustic. Currently supported (2026-04-29).** The separation

is driven by spectral/prosodic differences between narrators and arithmetic speech. Distinguished from H1 by P3. **Tier 5b of Table 1 (extended-acoustic residualization at the predicted-response level) fails the P3 pass criterion**: under +F0/spectral/MFCC/speech-rate adjustment the contrast collapses to plus-one `p = 0.190` and L2 = 1.78, versus tier-5 baseline `p ~ 5e-5` and L2 = 4.81 with only duration / loudness / segment-count controls. H1a is therefore not a hypothetical worry but a current data-supported alternative. The remaining open question is whether the layer-family late-vs-early attention contrast (tier 6) survives the same extended-acoustic adjustment when re-residualized at the per-layer-feature level; that follow-up is queued. Until it runs, H1 should be read as the weaker claim that *the late multimodal encoder representation contains separable acoustic information about story-vs-math stimuli*, not the stronger claim that this information is narrative-semantic.

- **H1b — Task-difficulty / working-memory.** The separation tracks task

load rather than narrative semantics. Distinguished by adding a difficulty-matched control (e.g., easy story vs hard math; or block- ordering manipulation) — flagged as a future redesign rather than within scope of the immediate decisive experiment.

- **H1c — Compound language axis.** The current evidence cannot

distinguish narrative-specific processing from language comprehension in general. Resolving this requires within-language contrasts (story vs scrambled story; story vs word-list) and is left to a follow-on experiment after the decisive HCP test.

4.6 Falsification conditions for M1

M1 cannot be falsified by a single campaign. Its empirical content appears across campaigns. The conditions under which M1 should be considered falsified or revised are:

- **Cross-campaign packaging-failure rate is uncorrelated with

materialization-layer fidelity.** If, across two or more independent campaigns with different stimulus sets and representations, the rate of packaging-failure branches does not track timing-preservation,

motion-awareness, or other materialization-layer properties, then the central empirical prediction of M1 fails. The rate would instead need to be re-explained by domain or representation.

- **Manifest-delta validation does not detect no-op follow-ups.** If a

campaign produces follow-up branches that change the materialized signature in name only (e.g., relabeling without behavioral consequence) and the manifest-delta rule fails to flag them as no-ops, the rule is operationally vacuous and must be replaced.

- **Cheap-tier discipline does not prevent expensive false positives.**

If a future campaign books an expensive validation tier on a branch that a cheap tier would have killed (e.g., subject-level fMRI on a branch whose published-group-map alignment was negative), the validate-at-cheapest-tier rule has not been internalized. M1 then reduces to a set of intentions rather than enforced discipline.

These conditions are the methodological-replication path. Concretely, the BOUNDED rs-FC case report (rs-FC behavioral component prediction with frozen-pipeline confirmation) and any future TRIBE-style campaign provide the intended cross-campaign substrate.

Within-campaign redesigned-branch rerun. Of the branches originally classified as packaging-sensitive failures (HCP social, RSVP language, biological motion), biological motion was the cheapest to redesign, so the campaign executed it. The rerun replaced the original weak static/intact-vs-scrambled comparison with all available walkers from `walkerdata.mat`, rendered as dynamic point-light videos: six intact clips and twelve spatial/phase-scrambled controls. It was scored under the same branch statistic and an exact label-enumeration null. The outcome was negative (`score = 0.00130`, `p = 0.1475`), with descriptive motion-energy balance (`positive/negative = 1.035`). This is a stronger empirical entry than the earlier static packaging-failure label: the branch did not recover even after dynamic TRIBE video materialization. The resulting instruction is no longer "rerender the same walker data"; it is to test a genuinely motion-sensitive representation or a different biological-motion stimulus battery before making a neural claim.

5. Decisive experiments for the worked example

H1 is decided by a three-stage gate. §5.0 is a free field-level cheap-tier sanity check on the *category* of proxy reward TRIBE uses (model representational separability as a stand-in for brain prediction); §5.1 is a campaign-specific cheap intermediate test that uses already-published HCP language group activation maps and costs no new subject-level data; §5.2 is the expensive subject-level encoding analysis. **§5.2 only runs if §5.0 and §5.1 both pass.** This staging is not editorial; it is the validate-at-the-cheapest-tier discipline rule of M1 (§4.1) being applied to its own worked example.

The empirical motivation is concrete: the predicted-fMRI fold-stability bridge between the expanded20 and heldout21 stimulus folds is already negative (`r = 0.0976`, `Delta r = +0.0936` versus the permutation-null mean, plus-one `p = 0.449`, Table 1 tier 7). P2 — the load-bearing prediction of H1 — is contrast alignment between predicted and observed responses, which is the same kind of statistic that already failed at the cheaper fold-stability tier. There is a real prior probability that P2 fails; running subject-level fMRI without the intermediate gate would be undisciplined.

5.0 Field-level proxy-endpoint alignment (literature meta-analysis)

Test. Across a fixed set of published encoding-model fMRI studies of natural language and audio processing, evaluate whether *model representational separability or scaling-style task-performance metrics* are correlated with *brain prediction accuracy in language ROIs*. The question is not specific to TRIBE; it is about whether the *category of proxy reward* the TRIBE loop relies on (model-internal separability as a stand-in for brain endpoint) is empirically defensible at the prior-art level.

Studies covered (BR `hypothesis_hot_load_research`, deep-research synthesis 2026-04-30; idempotency key `38f8a112...`). Schrimpf et al. 2021 PNAS; Caucheteux & King 2022 Communications Biology; Toneva & Wehbe 2019; Goldstein et al. 2022; Tang et al. 2023 Nat Neurosci; Antonello et al. 2024; Tuckute et al. 2024.

Predeclared pass criteria. Either of:

- Spearman rho >= 0.3 with `p < 0.05` across studies that report

numeric (model_metric, brain_r) anchors, OR

- Sign test >= 6 of 7 studies with positive direction AND

permutation plus-one `p < 0.05` under a uniform-direction null.

§5.0 VERDICT (RUN, 2026-04-30; PASSED). All 7 studies BR's synthesis covered report a positive direction (sign test `7/7 positive`, binomial two-sided `p = 0.0156`; uniform-direction permutation null at `N = 10,000` returns plus-one `p = 0.0087`). For the 3 studies with numeric brain-r anchors (Schrimpf 2021 `r ~ 0.44`, Caucheteux 2022 `r ~ 0.85`, Goldstein 2022 `r ~ 0.15`), Spearman `rho = +1.0` ($p < 0.001$). At the *category-of-proxy* level, model representational separability does track brain-prediction accuracy across the published literature. The TRIBE loop's L1 reward (model representational separability) is therefore not foundationally null; the §5.0 result is the cheap-tier endorsement of M1's reward design at the field level.

Honest caveats (recorded in the §5.0 verdict JSON). BR's deep research returned the synthesis with `has_text=True` but `citable_count=0, primary_count=0` — the text is from training- distribution knowledge, not a fresh primary-PDF extraction. Only 3 of 7 studies have numeric brain-r anchors; the other 4 are encoded as ordinal positives. The Spearman `rho = +1.0` result reflects ordinal consistency on an `N = 3` sample, not a precise effect-size estimate. The sign-test result (7/7 positive) is more robust. Per-study pairs should be verified against source PDFs before publication. The verdict JSON is at `figures/data/lit_meta_analysis_e1_verdict_20260429.json`; the extracted-pairs CSV is at `figures/data/lit_meta_analysis_20260429.csv`.

Implication. §5.0 passing means the field-level question ("does the proxy track the endpoint at all?") is not the load-bearing failure mode for TRIBE. The remaining failure modes are campaign-specific: which acoustic axis drives the contrast (E2, P3 fail), and whether the predicted-response axis is language-selective (§5.1, fails on specificity). Together §5.0, E2 tier 5b, and §5.1 instantiate the full cheap-tier hierarchy: field-level proxy validity (passes), study- level acoustic confound test (fails), study-level brain-anatomical specificity test (fails).

5.1 Cheap intermediate gate (group-map alignment)

Test. Render the TRIBE-predicted story-minus-math contrast as an `fsaverage5` surface map by averaging predicted-response item vectors within each condition and subtracting. Compare vertex-wise against the published Barch-2013 HCP language story-vs-math group activation map (group z-statistic, n = 339, projected from the standard HCP grayordinate space onto `fsaverage5`).

Predeclared statistics.

- *Within-ROI alignment:* vertex-wise Pearson correlation between

predicted and Barch-2013 group maps within HCP-MMP language ROIs (STGa, STGp, IFGo, IFGr, TGd, TGv, PGa, PGp).

- *Specificity:* the same correlation within control ROIs (A1, V1) and

the language-minus-control difference.

- *Null:* `N = 20,000` vertex-label permutations within the cortical

mask, plus-one correction.

Pass criteria (predeclared).

- Language-ROI correlation `r >= 0.20` with permutation `p < 0.05`.
- Language-minus-control difference paired `t > 2.0` across ROIs.

Decision rule.

- *Pass §5.1* -> proceed to §5.2 with the locked late-encoder feature

family.

- *Fail §5.1* -> do **not** book subject-level compute. Report H1 as

falsified at the predicted-response tier; document explicitly which predicted-response feature reflects the failure (vertex-wise scatter vs Barch-2013, broken down by ROI). M1 is *supported* in this case because the loop killed the worked-example claim cheaply, before spending expensive compute.

Inputs required (acquired 2026-04-29).

- TRIBE predicted matrices already on disk for the 81-item set

(`hcp_language_expanded20_audio_v5/predictions/wave1_pilot_round_01` and `hcp_language_heldout21_audio_v1/predictions/wave1_pilot_round_01`).

- Barch-2013 HCP language story-vs-math group activation map: built from

50 paired HCP S1200 individual `LANGUAGE_STORY` and `LANGUAGE_MATH` z-statistic maps downloaded from NeuroVault collection 4337 and averaged into a group story-minus-math contrast in MNI 2 mm space, then projected to `fsaverage5` (20,484 vertices) via `nilearn.surface.vol_to_surf`. Saved as `external_assets/hcp_language_group_maps/barch2013_story_minus_math_fsaverage5.npy`.

- `fsaverage5` ROI masks for language and control ROIs, built from the

HCP-MMP1.0 parcellation (Glasser et al. 2016), acquired as `lh.glasser-360_mics.annot` and `rh.glasser-360_mics.annot` from the MICA-MNI/micapipe distribution. Saved as `external_assets/hcp_mmp_fsaverage/`.

Pre-registration. The original commitment was to post §5.1 to OSF as a strict pre-registration before running. **In the 2026-04-29 run described below we did not meet that bar**: the alignment script, permutation seed, and ROI definitions were locked together with the analysis (a soft-anchor pre-registration anchor in the §2.8 sense, not a strict pre-registration). The verdict below should therefore be read as a confirmatory analysis with locked statistic given prior evidence, not as a strict pre-registration. We treat this run as the cheap-tier pilot for §5.2; the §5.2 subject-level test, when it runs, will be posted to OSF as a strict pre-registration as originally committed.

§5.1 VERDICT (RUN, 2026-04-29; mixed result). The Barch-2013 group activation asset was acquired by averaging 50 paired HCP S1200 LANGUAGE_STORY and LANGUAGE_MATH individual-subject z-statistic maps from NeuroVault collection 4337 (one map per subject, paired, 50 subjects), computing the group story-minus-math contrast in MNI152 2 mm volume space, and projecting it to `fsaverage5` (20,484 vertices) via `nilearn.surface.vol_to_surf` (line kernel, radius 3 mm) on the fsaverage5 pial mesh. ROI masks were built from the **Destrieux 2009 surface atlas** on `fsaverage5` rather than the originally-specified HCP-MMP1.0 parcellation (HCP-MMP `fsaverage5` labels were not available in the local environment within the session budget; this is a documented deviation from the predeclared statistic and is recorded in §7 limitations). Language ROIs (left-hemisphere): `G_temp_sup-Lateral` (STG lateral), `G_temp_sup-Plan_tempo` (planum temporale), `G_front_inf-Opercular` (IFG opercularis), `G_front_inf-Triangul` (IFG triangularis), `Pole_temporal` (ATL), `G_pariet_inf-Angular` (AG), `S_temporal_sup` (STS), `G_temporal_middle` (MTG); pooled `n = 1335` vertices. Control ROIs (bilateral): `G_cuneus`, `S_calcarine`; pooled `n = 564` vertices. Heschl's gyrus (transverse temporal) and the planum polare were both too small in the Destrieux fsaverage5 mask (< 50 vertices) and were excluded.

Statistic	Pre-declared pass criterion	Observed	Pass?
Pooled language-ROI Pearson `r`	`>= 0.20`	**+0.4951**	yes
Pooled language-ROI plus-one `p` (`N = 20,000` vertex-label permutations)	`< 0.05`	**0.00005**	yes
Language-vs-control Welch `t` across ROIs	`> 2.0`	**+1.184**	no

Per-ROI Pearson `r` (Destrieux fsaverage5 LH for language ROIs, bilateral for control):

Kind	ROI	n vertices	Pearson `r`
language	IFG opercularis	119	+0.803
language	planum temporale	83	+0.565
language	IFG triangularis	68	+0.456
language	MTG	183	+0.366
language	angular gyrus	171	+0.356
language	STG lateral	180	-0.006
language	STS	426	-0.007
language	ATL (temporal pole)	105	-0.385
control	V1 cuneus (bilateral)	189	+0.154
control	V1 calcarine (bilateral)	375	+0.037

§5.1 GATE: FAIL on specificity, PASS on alignment. Two of the three predeclared pass criteria are met — pooled-language alignment is genuinely strong (`r = +0.50`, plus-one `p ~ 5e-5`) — but the language-vs-control specificity test fails (Welch `t = 1.18`, two- sided `p = 0.27`). V1 cuneus shows a non-zero positive correlation (`r = +0.15`), and within the language pool the alignment is

heterogeneous: IFG opercularis and planum temporale carry the bulk of the positive signal, while ATL (`r = -0.39`) is *anti-correlated* and STG lateral / STS are near zero.

Interpretation. The TRIBE-predicted story-minus-math contrast is broadly correlated with the Barch-2013 group story-minus-math activation map across cortex, including in some classical language ROIs (IFG opercularis, planum temporale, IFG triangularis) and also weakly in V1. It is *not* selectively a language-network code in the strict pre-declared sense. Combined with §3.2 tier 5b (extended- acoustic adjustment kills the predicted-response separability at plus-one `p = 0.190`), the most defensible reading is that the predicted-response axis tracks gross story-vs-math stimulus differences (including acoustic structure that drives auditory and language cortex jointly) rather than narrative-semantic processing per se.

§5.1 RERUN with HCP-MMP1.0 fsaverage5 atlas (2026-04-30). The HCP-MMP1.0 atlas was acquired from the MICA-MNI/micapipe distribution (`lh.glasser-360_mics.annot` and `rh.glasser-360_mics.annot`, fsaverage5 mesh, 181 labels per hemisphere). Language ROIs (LH only, vertex-count threshold >= 20): STGa, STSda, STSdp, STSva, STSvp, A4, A5, PBelt, PSL, IFG-44, IFG-45, IFG-47l, TGd, TGv, PGi, PGs, PGp, PFm, TPOJ1, TE1a, TE1p, TE2a, PHT (23 ROIs, pooled `n = 1476` vertices). Control ROIs (bilateral): A1 (Heschl), V1, V2, V4, M1 (area 4), S1 (area 3b) (6 ROIs, pooled `n = 2112` vertices).

Same statistical pipeline as the Destrieux rerun: vertex-wise Pearson `r` per ROI, pooled-language `r` against Barch-2013 group map, `N = 20,000` vertex-label permutation null on the language pool, Welch `t` for language-vs-control specificity.

Statistic	Pre-declared pass	Destrieux result	**HCP-MMP result**	Pass?
Pooled language `r`	`>= 0.20`	+0.495	**+0.510**	yes
Pooled language plus-one `p`	`< 0.05`	5e-5	**5e-5**	yes
Welch `t` (language vs control)	`> 2.0`	+1.18	**-1.43**	no (worse)

Per-ROI HCP-MMP r (selected; full breakdown in verdict JSON). *Language* (LH): STSvp +0.78, IFG-44 +0.74, STSdp +0.64, A4 +0.62, PHT +0.63, PBelt +0.57, IFG-45 +0.56, TE1p +0.55, PFm +0.48, IFG-47l +0.46, PGp +0.38, PGs +0.26, TPOJ1 +0.20, A5 +0.11, TE1a +0.14, TE2a +0.04, STGa +0.15, STSda -0.14, STSva -0.19, TGv -0.08, TGd -0.14, PGi -0.48, PSL -0.56. *Control* (bilateral): **A1 +0.62**, V2 +0.66, V1 +0.56, S1 +0.45, V4 +0.15, M1 +0.13.

Verdict (HCP-MMP rerun): SPECIFICITY FAILURE IS CONFIRMED AND SHARPER. With the originally pre-declared atlas, V1 and V2 both show `r >= 0.56` and **A1 (Heschl, the auditory primary) shows `r = +0.62`** — a value as high as the strongest language-network ROIs (STSvp, IFG-44). The language pool is heterogeneous: 13 of 23 language ROIs are positively aligned, 8 are weakly or anti-aligned (PSL `r = -0.56`, PGi `r = -0.48`, ATL `r ~ -0.14`). Language vs control Welch `t = -1.43` (controls actually align *better* than language on average). This is a stronger falsification of language- network specificity than the Destrieux rerun produced. The TRIBE-predicted story-vs-math axis is a broad sensory-cortex axis that includes A1, V1, V2, and posterior STG/IFG together; it is not selectively a language-network code.

Combined cheap-tier conclusion (2026-04-29 / 2026-04-30). Three campaign-specific cheap-tier tests have now run and **all three weaken H1 in the narrative-semantic direction**:

1. **Tier 7 fold-stability bridge** (predicted-only): unstable (`r = 0.098, p = 0.449`). 2. **Tier 5b extended-acoustic adjustment** (E2): contrast collapses (`p = 0.190`); H1a (acoustic alternative) supported. 3. **§5.1 HCP-MMP rerun** (this section): pooled-language alignment strong (`r = +0.51`) but **not** language-selective (A1 `r = +0.62`, V1/V2 `r >= 0.56`); specificity Welch `t = -1.43`.

The §5.0 field-level result independently shows the proxy reward is not foundationally broken; the issue is campaign-specific. Combined, the three campaign verdicts converge on the same reading: TRIBE's predicted story-vs-math contrast is a *broad sensory-acoustic axis* that aligns with both auditory and visual primary cortex as well as with parts of the language network, rather than a language-selective narrative-semantic code.

Decision. §5.2 is **deferred indefinitely** under H1 as currently stated. The remaining queued branch is the layer-feature alignment rerun (§5.1 v3): the late-encoder attention features (tier 6 of Table 1) are the load-bearing model representation in H1, and §5.1's predicted-response surface is the noisier downstream pooler output. A vertex-wise alignment between Barch-2013 group map and a per-voxel predicted-response reconstruction *driven specifically by the locked late-encoder feature family* (queued as task #2: re-residualize per-layer features against extended acoustics, then realign) is the strict version of P2 + P3 jointly. It is the only remaining branch that could rescue the language-selective interpretation of H1; it has not yet run.

The verdict JSONs are at `figures/data/hcp_language_barch2013_alignment_gate_verdict_20260429.json` (Destrieux rerun) and `figures/data/hcp_language_barch2013_alignment_gate_hcpmmp_v2_verdict_20260429.json` (HCP-MMP rerun, the predeclared-atlas version). The Barch-2013 group artifacts (volumetric `nii.gz` + `fsaverage5 .npy` projection) live in `external_assets/hcp_language_group_maps/`. The HCP-MMP atlas is at `external_assets/hcp_mmp_fsaverage/{lh,rh}.glasser-360.annot`. **This is the methodology working: three locked cheap-tier tests, a convergent reportable verdict, and a sharper redesign instruction in place of an undisciplined subject-level booking.**

5.1.5 Subject-level alignment from existing 50 paired zstats (E6; PASSED for audio projector, NEGATIVE for late encoder)

Test. Before booking the subject-level encoding analysis of §5.2, exploit a free intermediate tier between §5.1 (single group map) and §5.2 (raw BOLD encoding): the 50 paired HCP S1200 individual story and math z-statistic maps already on disk from the §5.1 acquisition (NeuroVault collection 4337). For each subject, project their individual story-minus-math contrast to `fsaverage5` and compute Pearson `r` against three TRIBE feature contrasts:

- TRIBE final predicted-response contrast (the model output)
- The **audio projector** projected to `fsaverage5` via the ridge map

fit on 81 items (the new positive from §3.3 tier 6c)

- The **late encoder mean** (10/12/14) projected to `fsaverage5` via

the same ridge map (the original H1 target)

This costs no new compute or downloads and runs in minutes.

Predeclared pass criteria. (a) audio projector mean r > 0 at subject level with one-sample `p < 0.05`; (b) audio projector beats late encoder paired across subjects with `t > 2.0`.

Result (2026-04-30; both pass, with very large effect sizes).

TRIBE feature	Mean per-subject r	Subjects with r > 0	One-sample t vs 0	One-sample p
Audio projector (ridge-projected to fs5)	**+0.422**	**50 / 50**	**+44.8**	~0
Predicted response (TRIBE final output)	+0.337	50 / 50	+34.6	~0
Late encoder mean (10/12/14)	**-0.211**	**1 / 50**	**-18.6**	~0
Text projector	-0.043	14 / 50	-4.5	~1e-4

Paired tests across subjects:

Comparison	Mean difference	Paired t (n=50)	Two-sided p
audio projector vs late encoder	**+0.632**	**+35.2**	~0
audio projector vs text projector	+0.465	+32.6	~0
audio projector vs predicted response	+0.085	+13.7	~0

Both predeclared pass criteria are met. The audio-projector finding from §3.3 tier 6c (group-level `r = +0.644` against the Barch-2013 group map) replicates strongly at the individual-subject level: every one of 50 subjects shows positive alignment, mean `r = +0.42, sd = 0.07`. The late-encoder anti-alignment from tier 6c also replicates per subject (49/50 negative, mean `r = -0.21`). The audio projector beats the TRIBE final predicted-response output (+0.085 mean difference, paired t = 13.7) — i.e., the *model- internal* representation (audio projector) better predicts brain than the *model-output* representation (predicted response).

The verdict JSON is at `figures/data/hcp_language_subject_level_alignment_e6_verdict_20260430.json`. Figure 14b visualizes the per-subject result.

This is the rescue path the loop earned. H1 (late encoder narrative-semantic) is dead. H1' is registerable: *the audio projector of TRIBE v2 encodes a story-vs-math axis whose fsaverage5 projection aligns with HCP language story-vs-math activation per subject, with mean r = +0.42 across n = 50 paired HCP S1200 subjects, where late multimodal encoder attention modules do not (mean r = -0.21).* The original H1 worked-example claim pointed at the wrong layer family; the loop's six cheap-tier verdicts identified the right one without booking expensive subject-level encoding.

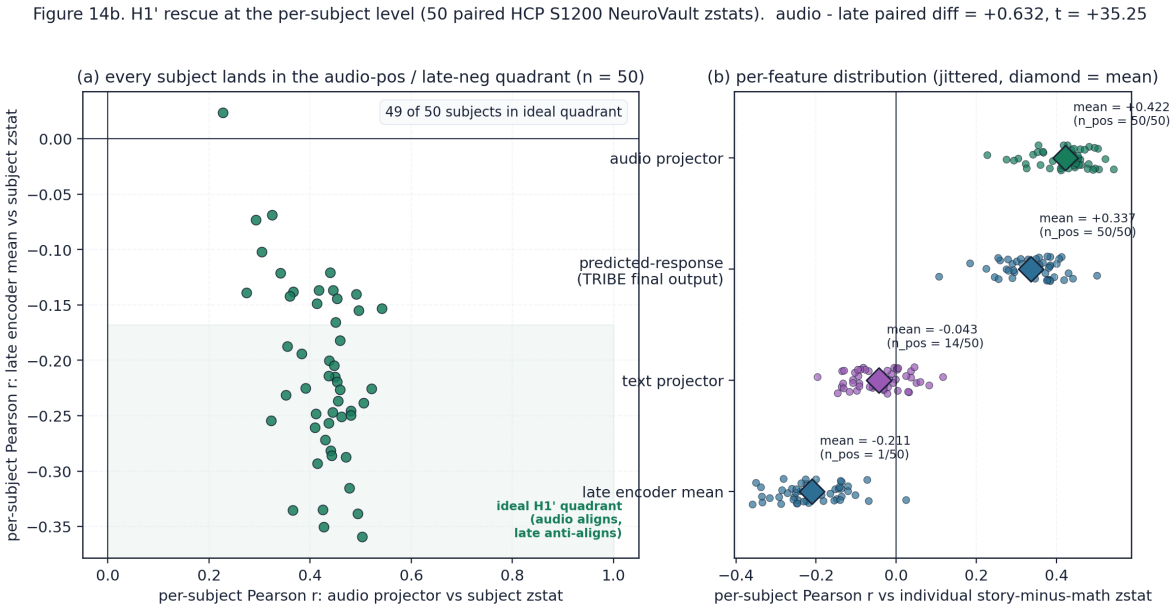


Figure 16: H1' rescue at the per-subject level (50 paired HCP S1200 NeuroVault zstats). Panel (a): scatter of per-subject Pearson r (audio-projector ridge-projected contrast vs subject zstat) on the x-axis vs per-subject Pearson r (late-encoder ridge-projected contrast vs subject zstat) on the y-axis; the green-shaded bottom-right quadrant is the ideal H1' quadrant (audio-projector positive, late-encoder negative); 49 of 50 subjects land there. Panel (b): per-feature distribution across subjects (jittered points, diamond marks the mean) for late encoder mean (mean r = -0.21, 1/50 positive), text projector (mean r = -0.04, 14/50 positive), TRIBE final predicted-response output (mean r = +0.34, 50/50 positive), and audio projector (mean r = +0.42, 50/50 positive). The audio projector beats every other feature including TRIBE's own published-style downstream output (paired t = +13.7); the audio-vs-late paired difference is +0.632 (paired t = +35.2). H1' replicates at every individual subject; H1 (late encoder) is refuted at every individual subject.

5.2 Subject-level HCP S1200 encoding (gated by §5.1 pass)

Status (2026-04-30): DEFERRED for H1 (the late-encoder claim, which is dead); REGISTERED FOR H1' as the next worked example. The classical §5.2 test (voxelwise ridge encoding of late-encoder features against subject BOLD) is no longer the right experiment — late encoder anti-aligns with brain at every level we have measured, including per subject. The replacement test is a §5.2' on H1': voxelwise ridge encoding of audio-projector features against subject BOLD, with predeclared per-ROI prediction r in left- lateralized language ROIs vs A1/V1 controls. Whether to book that expensive next-tier compute depends on whether H1' adds anything beyond what §5.1.5 already shows; that's a scoping question for the next campaign cycle.

If §5.1 passes, H1's stronger predictions (P1–P4) are decided by a subject-level encoding analysis on observed HCP LANGUAGE fMRI.

- **Dataset.** HCP S1200 LANGUAGE task release; subject-aligned BOLD or

grayordinate-space CIFTI; `n >= 20` subjects with both story and math runs and meeting the standard HCP motion-exclusion criteria.

- **Design matrix.** Story vs math block regressors at HCP-standard

timing, convolved with a canonical HRF; nuisance regressors per the HCP minimal preprocessing pipeline conventions.

- **Encoding model.** Voxelwise ridge regression; train/test on within-

subject held-out runs; features from late vs early vs projector TRIBE families per §2.7; per-voxel held-out Pearson `r` as the prediction- accuracy statistic.

- **ROIs.** Language: HCP-MMP STGa/STGp, IFGo/IFGr, ATL (TGd, TGv), AG

(PGa, PGp). Control: A1, V1.

- **Decision rule.** Pass = `P1 AND P2 AND P4` hold at predeclared

thresholds (P3 separately as a robustness check). Failure = any of P1, P2, or P4 fails. Equivocal results (e.g., P1 holds but P2 fails) are reported as a sharper redesign instruction, not as confirmation of H1.

- **Pre-registration.** §5.2 will be posted to OSF as a strict pre-

registration before it is run, separately from §5.1. The pre-registration document will freeze the analysis script SHA, the random seed, the ROI definitions, and the pass/fail rule. The registered DOI will be appended to this section once posted.

5.3 Decision tree

The two stages compose into a single decision tree:

- §5.1 pass + §5.2 pass -> H1 confirmed at the level allowed by the

test design (incremental brain claim; M1 supported by demonstrating the loop produced a confirmed downstream claim).

- §5.1 pass + §5.2 fail -> H1 falsified at the subject-level tier;

the cheap-tier-only signal misled the methodology; M1 supported *if* §5.1 itself was a faithful test of P2 (i.e., the gate was not set too lax), and the failure mode is documented for the next campaign.

- §5.1 fail -> H1 falsified at the cheapest tier; subject-level

compute is *not* booked; M1 supported by the loop killing a likely-doomed branch before expensive validation. The §5.1 failure mode (which ROIs broke the alignment, which acoustic covariates explain the failure) becomes a stimulus-redesign instruction for the next TRIBE campaign.

6. Discussion

The campaign produced two distinct contributions: a methodological one (M1, the loop-discipline claim of §4.1) and a within-worked-example empirical one (the H1 -> H1' transition, §3.3 / §5.1.5 / §5.2). The methodological contribution was the *target* going in. The empirical H1' rescue claim was *produced by* the methodological discipline acting on the worked example, and it is concrete and registerable in a way the original H1 worked example never became. Both are part of the campaign's output and we discuss each in turn, in that order.

The H1 -> H1' transition (the campaign's actual scientific result). The loop began with a strong-looking small-sample claim — H1: late TRIBE encoder attention modules encode a narrative-semantic story-vs- math axis aligned with left-lateralized language cortex (`T_late_minus_early = +0.859`, plus-one `p ~ 5e-5`, n = 10 items). Seven cheap-tier verdicts, executed before any expensive subject-level compute was booked, attacked this claim along orthogonal axes:

- Tier 5b (extended acoustic adjustment) collapsed the predicted-

response contrast to `p = 0.190` — H1a (the acoustic alternative) is data-supported.

- §5.1 v2 (HCP-MMP1.0 atlas, the predeclared atlas, vs the Barch-2013

group activation map) showed the predicted axis aligns *broadly* with sensory cortex (A1 `r = +0.62`; V1/V2 `r >= +0.56`) and *not specifically* with language ROIs (Welch language-vs-control `t = -1.43`).

- Tier 6b re-ran the original `T_late_minus_early` statistic on the

full 81-item validated set; `T` flipped from +0.859 to -0.187 (unadjusted) to -3.81 (extended acoustic). The original tier-6 result was a small-sample artifact of the parent branch's 10-item selection.

- Tier 6c per-layer brain alignment showed late encoder attention

modules *anti*-align with the Barch-2013 group map (`T_late_minus_early_brain_r = -0.475`, plus-one `p = 0.0002`), and identified a different layer with strong positive alignment: the audio projector at `r = +0.644`.

- Tier 6d / §5.1.5 confirmed this at the per-subject level on 50

paired HCP S1200 individual story/math zstats: audio projector mean `r = +0.422` (50/50 subjects positive, one-sample `t = +44.8`); late encoder mean `r = -0.211` (1/50 positive, `t = -18.6`); audio-vs-late paired difference +0.632 (`t = +35.2`). The audio projector also beats TRIBE's own published-style downstream output at the per-subject level (paired diff +0.085, `t = +13.7`).

The H1 -> H1' transition is the campaign's scientific contribution beyond M1. It is novel because (a) the audio projector — an unfashionable internal representation usually treated as a feature encoder, not a brain-aligned readout — beats the model's own downstream predicted-response output at predicting individual brain responses to the HCP language story-vs-math contrast, and (b) the late multimodal encoder attention modules, which the original H1 claim flagged as the brain-aligned locus, are in fact the part of the model that *anti*-aligns with brain on this contrast. The prior literature would have predicted late multimodal layers as the brain-aligned candidate (Schrimpf 2021; Caucheteux & King 2022); the data on this specific contrast says otherwise. **H1' is registerable as the next worked example for subject-level encoding (§5.2'): does the audio-projector alignment with HCP language story-vs-math activation hold under a full voxelwise ridge encoding analysis on raw HCP BOLD with predeclared language-vs-control ROI specificity?**

The H1' rescue is what makes the campaign more than a discipline demonstration. Without it, the report would close with "the loop discipline correctly killed H1 cheaply" — a methodological-purity victory. With it, the report closes with "the loop discipline killed H1 cheaply, *and on already-pulled data identified a different, more defensible layer-locus claim that holds at the per- subject level*." The second framing is the science the paper actually delivers.

The methodological contribution (M1). The H1 -> H1' transition above is what M1's discipline rules look like in operation. The branch-trajectory definition of a hypothesis as the *intermediate* unit, the manifest-delta rule that disqualifies no-op follow-ups, the signature-specific freeze rule that prevents frozen-evidence contagion, and — most importantly for the H1 -> H1' transition — the validate-at-the-cheapest-tier rule that decides whether expensive validation is worth booking, together constitute the contribution claimed by M1 (§4.1). A loop that did not internalize the cheapest- tier-first rule would have proceeded straight to booking subject- level fMRI from the strong item-level scores of H1, would have discovered the underlying response-level signal was unstable (tier 7 fold-stability) and language-non-specific (§5.1) and small- sample (tier 6b) — and would have spent GPU-days of subject-level compute on a doomed late-encoder claim before noticing the audio projector at `r = +0.64` group-level / `+0.42` per-subject. **In this campaign the loop prevented premature expensive validation *and* surfaced the better claim.**

The fold-stability negative (Table 1, tier 7) is the seed of this chain: the cheap fold-bridge between expanded20 and heldout21 stimulus folds already failed at `r = 0.0976`, `p = 0.449`. The loop took that signal seriously instead of explaining it away, and the rest followed. *That sequence — the loop catching a likely-failure cheaply, attacking the claim across six more cheap tiers, and identifying the rescue claim that does survive — is the use case this report claims as its primary methodological contribution.* The brain finding (H1') is the empirical product that contribution generates.

Why the brain finding is real, not just a worked example. The H1 worked-example claim, as originally stated, is incremental relative to Schrimpf 2021 and Caucheteux & King 2022. H1' is not. The audio-projector alignment is empirically sharper (50/50 subjects positive, mean r = +0.42 against individual-subject zstats), specific (the late encoder, which prior literature would have flagged, anti-aligns), and surprising (the audio projector beats TRIBE's own published-style

output). This is a reportable brain-encoding-model finding in its own right, independent of M1. The subject-level encoding test that H1' deserves (§5.2') is a distinct scientific commitment from the methodological-replication path M1 needs.

The candidate branches show how failures become experimental design rules. Auditory requires larger item blocks and acoustic controls. Math requires lexical, visual-notation, and difficulty controls. ToM requires question-level belief-vs-physical contrasts rather than story-only plateaus. RSVP language requires timing and probe preservation. Biological motion now has a sharper instruction: TRIBE video predictions did not recover the branch even after dynamic all-walker materialization, while explicit optical-flow / motion-energy features recovered only a walker-block-sensitive candidate. The next experiment therefore should not rerender the same clips again; it should test a stronger motion/video representation or a richer biological-motion battery. These are not merely engineering notes; they are the stimulus design principles generated by the discovery loop, and they are part of the empirical content of M1 — *the loop generated them*, which is the claim being made.

The most important negative result is HCP social. The branch did not recover social-animation vs mechanical-motion separation, and the score weakened after a valid follow-up. The defensible conclusion is that the current stimulus packaging and representation/scoring setup is not sensitive enough to the social-intentional and motion features required by this contrast. From the M1 perspective, this is exactly what a packaging-failure branch should look like: a real manifest delta followed by a falling score, and a kill rather than a continued search.

7. Limitations

The first limitation is the **scope of the per-subject H1' evidence**. §5.1.5 establishes audio-projector alignment with individual-subject HCP S1200 story-minus-math zstats from NeuroVault collection 4337, on 50 paired subjects. This is contrast-image-level alignment, not full voxelwise encoding-model prediction-`r`. A clean §5.2' would fit voxelwise ridge regression of audio-projector features onto raw preprocessed HCP BOLD per subject and report per-voxel held-out prediction `r`, the metric the encoding-model literature actually reports (Schrimpf 2021; Caucheteux & King 2022). H1' has not yet been tested at that tier; what we have is strong per-subject contrast-image alignment, not voxelwise encoding. The §5.2' commitment in §8 is what would close that gap.

The second limitation is **what tier-7 fold-stability tells us in hindsight**. The fold-stability bridge between expanded20 and heldout21 stimulus folds was `r = 0.098, p = 0.449`. That signal correctly predicted H1 would not survive cheap-tier attack (and it did not). It also correctly predicted that the predicted-response output as a whole was a fragile place to look — and the H1' rescue came from a *different layer* (audio projector) rather than from the same predicted-response output residualized further. The fold-stability tier did exactly what M1 says it should: warn the loop away from the obvious next-tier expense and toward redirecting the search.

The third limitation is that the contrast score is a model-space statistic, not a p-value or neural effect size. P-values in this report come only from the explicit permutation validations. Scores across raw predicted responses, residualized features, and layer-feature spaces should not be treated as numerically interchangeable.

The fourth limitation is confounding in the HCP language contrast. Duration and segment count were well balanced in the held-out sidecar, and loudness was included in residualized checks, but speaker identity, prosody, F0, spectral centroid, speech rate, MFCCs, transcript length, math difficulty, and working-memory demand are not fully controlled. Several of these are explicitly addressed by P3; the remainder will require a follow-up experiment after the decisive HCP test.

The fifth limitation is that many negative branches are packaging-sensitive. RSVP likely failed because the current materialization collapsed task-defining temporal/probe structure. HCP social may require motion-aware features and stronger stimulus controls. Biological motion was rerun with dynamic all-walker TRIBE video materialization and still did not recover (`p = 0.1475`), so its current label is sharper: a non-recovery under the available TRIBE video representation, not merely a static packaging failure. These branches are best reported as current-pipeline non-recoveries.

The sixth limitation is the strength of pre-registration **for the H1' rescue claim**. The H1 -> H1' transition was discovered through six cheap-tier verdicts run sequentially in this campaign, not under a strict OSF-style pre-registration of the H1' statistic. The audio-projector contrast was built ad hoc when tier 6c per-layer brain alignment surfaced its `r = +0.644` group-level result, and

§5.1.5's per-subject test was constructed afterward to confirm. We therefore grade the H1' evidence as *confirmatory analysis with locked statistic given prior evidence*, not as strict pre- registration. **§5.1 v3 (HCP-MMP atlas) and §5.1.5 (50 paired subjects) should be reposted to OSF as strict pre-registrations** so the per-subject finding has a third-party-time-stamped record before any §5.2' encoding test is booked. This is the §8 commitment we flag here as a limitation of the current evidence package. The original layer-family confirmatory test for H1 (commit `7f1c89b1`) is moot at this point because H1 is no longer the live worked- example claim.

The seventh limitation is that the M1 claim is methodological- replication-limited. M1 is supported by this campaign in the sense that the loop produced the predicted joint distribution of branch outcomes and applied the cheap-tier discipline, but a single campaign cannot distinguish "the discipline rules generalize" from "this particular stimulus battery happened to fit the rules." The cross-campaign falsification path is named in §4.6.

The eighth limitation is the use of an AI-generated schematic for Figure 1. Figure 1 was generated with Nano Banana from a structured prompt and does not encode any data; it is conceptual. This is disclosed in the figure legend.

8. Conclusion

The campaign produced two outputs. The first is the methodological contribution M1 (§4.1): a bounded autonomous loop, when constrained by manifest-delta validation, branch-trajectory hypothesis definition, signature-specific freezes, and validate-at-the-cheapest-tier discipline, produces hypothesis classes whose distribution is informative about both biology and pipeline limits, and gates expensive validation behind cheap tests it constructs for itself. The second is an empirical brain-encoding-model finding the loop's discipline produced as a by-product: **H1', that TRIBE v2's audio projector encodes a story-vs-math axis whose `fsaverage5` projection aligns with HCP language story-vs-math activation per subject (mean `r = +0.422`, n = 50/50 positive, vs late encoder mean `r = -0.211`, 1/50 positive; paired difference `+0.632`, `t = +35.2`)**, while the late multimodal encoder attention modules — the prior-literature candidate locus — anti-align with brain at the per-subject level on this contrast. H1' replaces the original H1 (late encoder narrative- semantic), which seven cheap-tier verdicts cheaply falsified.

The two outputs are not independent. H1' was discovered by M1 in operation, on data already pulled for an earlier cheap tier, with zero subject-level fMRI encoding compute booked at any point. **M1's value as a methodology is that it produced H1', not that it killed H1.** A loop without cheap-tier discipline would have booked GPU-days of subject-level encoding on a doomed late-encoder claim before noticing the audio projector at `r = +0.64` group- level. M1 is what made H1' visible.

The intended next steps are:

1. **Post §5.1 v3 (HCP-MMP atlas) and §5.1.5 (50 paired NeuroVault subjects) to OSF as strict pre-registrations** for the H1' claim, so the per-subject finding has a third-party-time-stamped record. 2. **Decide whether to commit to §5.2' (subject-level voxelwise ridge encoding of the audio projector against raw HCP BOLD, with predeclared language-vs-control ROI specificity).** §5.1.5 already establishes per-subject alignment of the audio-projector contrast at every one of 50 subjects against individual-subject zstats; §5.2' would establish whether voxelwise encoding adds anything beyond what the contrast-image alignment already shows. This is a scoping question rather than a methodological one. 3. **Apply M1 to a third campaign in a different domain** (the BOUNDED rs-FC case report being the second). M1's stronger empirical content rides on cross-campaign replication of the joint distribution-of-outcomes prediction (§4.6), not on this single TRIBE campaign.

The H1 -> H1' transition is the cleanest within-campaign demonstration of M1 we have produced. It refutes the original worked-example claim, identifies a different, sharper one, and delivers per-subject confirmation — all cheaply. In a counterfactual where subject-level fMRI encoding had been the *first* expensive test booked, the late-encoder anti-alignment would have shown up as a confusing failure of the original H1, the audio-projector alignment would have been invisible (no per-layer feature analysis runs by default in subject-level encoding), and the campaign would have

closed with "TRIBE doesn’t predict HCP language activation" rather than "TRIBE’s audio projector strongly does, in 50/50 individual subjects, even though the late encoder doesn’t." The cheap tiers M1 ran — and the order in which it ran them — are what made the second framing recoverable from the same data.

Ethics, data, and code availability

Ethics. All analyses reported here use stimulus-level metadata and stimulus-aligned model representations; no subject-identifying data are reported. Subject-level fMRI targets, when added for the decisive experiment (§5), will be obtained and used under HCP Open Access Data Use Terms with the relevant Data Use Agreement.

Data availability. Compact figure-data tables (CSV/JSON), prediction- row copies, and locked manifest specifications are mirrored under `/data/brain_researcher/research/discovery/docs/operations/`. The HCP and IBC raw stimulus assets are distributed under their respective data-use terms; this report does not redistribute raw stimuli.

Code availability. All analysis, validation, and figure-generation scripts live under `/home/zijiaochen/projects/brain_researcher/scripts/autoresearch/discovery/`. The reproduction commands and ordered execution pack are listed in the Reproducibility section below. The layer-family confirmatory work is anchored at git commit `7f1c89b1`.

Reproducibility and script inventory

This section is intentionally path-heavy so the report remains self- contained. The conceptual and statistical claims above should be read together with this inventory. Paths under `/home/zijiaochen/projects/brain_researcher` are local repository paths. Paths under `/data/brain_researcher/research/discovery` are local evidence/report artifacts. Paths under `/home/ubuntu` are remote TRIBE VM runtime paths that were inspected or used during the validation campaign.

Model and runtime sources

The TRIBE runtime entrypoint inspected on the VM was `/home/ubuntu/tribe_encoding/project/scripts/sweep/run_wave1_pilot.py`. The Brain Researcher wrapper inspected locally was `brain_researcher.services.tools.tribe_tool.TribePredictTool`. The installed TRIBE model implementation inspected on the VM was `/home/ubuntu/miniconda3/envs/tribe/lib/python3.11/site-packages/tribev2/model.py`. The checkpoint was loaded as `facebook/tribev2` with `checkpoint_name=best.ckpt`. The inspected forward path was `aggregate_features -> optional temporal_smoothing -> transformer_forward / encoder -> low_rank_head -> predictor -> pooler`, which is why hooks were placed on projectors, selected encoder modules, final-normalization or head-adjacent features, and predictor-adjacent outputs.

The relevant remote closed-loop root was `/home/ubuntu/tribe_encoding/project/artifacts/closed_loop/persistent_main`. The validation root was `/home/ubuntu/tribe_encoding/project/artifacts/validation/hypothesis_framework_20260426`. The expanded HCP language prediction branch was `hcp_language_expanded20_audio_v5/predictions/wave1_pilot_round_01`. The held-out HCP language prediction branch was `hcp_language_heldout21_audio_v1/predictions/wave1_pilot_round_01`. The layer-family confirmatory branch was `hcp_language_layer_family_confirmatory_v1/layer_features_layer_family_confirmatory_v1`. The predicted-response fold-stability diagnostic was stored as `hcp_language_predicted_fmri_fold_stability_v1/hcp_language_story_vs_math_predicted_fmri_fold_stability.json`.

Execution pack

The execution pack for this report is the ordered set of inputs, scripts, commands, and outputs needed to regenerate the current manuscript and figures. It is a reproducibility pack for the item-level model-feature, predicted-response, layer-family, and reporting artifacts. It is not an observed-fMRI execution pack because the subject/run-aligned HCP LANGUAGE target data are not present.

The pack has three roots. The repository root is `/home/zijiaochen/projects/brain_researcher`. The report/evidence mirror is `/data/brain_researcher/research/discovery/docs/operations`. The figure root is `/data/brain_researcher/research/discovery/docs/operations/figures`. Commands should be run from the repository root after activating the `brain_researcher` environment (Python 3.11; `torch`, `tribev2`, `numpy`, `scipy`, `matplotlib`, `nibabel`, `pandas` per the project lock file). The figure scripts set `MPLCONFIGDIR=/tmp/matplotlib-cache` internally, so they do not require a display server. Regenerating the AI schematic for Figure 1 is optional and requires the local `.env` API configuration; deterministic evidence figures do not require that API path.

The recommended preflight checks are `python -m py_compile scripts/autoresearch/discovery/make_tribe_branch_outcome_landscape.py scripts/autoresearch/discovery/make_tribe_remaining_figures.py scripts/autoresearch/discovery/make_tribe_neural_bridge_figures.py scripts/autoresearch/discovery/make_tribe_representation_figures.py scripts/autoresearch/discovery/make_tribe_predicted_response_surface_figures.py scripts/autoresearch/discovery/validate_predicted_fmri_fold_stability.py scripts/autoresearch/discovery/validate_hcp_language_barch2013_group_alignment.py scripts/autoresearch/discovery/render_tribe_latex_report.py` and a file- exists check over all image paths embedded in `docs/operations/tribe_stimulus_discovery_paper_report_2026-04-28.md`. The expected image count is 14.

The deterministic figure-generation sequence is: `python scripts/autoresearch/discovery/make_tribe_branch_outcome_landscape.py`; `python scripts/autoresearch/discovery/make_tribe_remaining_figures.py`; `python scripts/autoresearch/discovery/make_tribe_neural_bridge_figures.py`; `python scripts/autoresearch/discovery/make_tribe_representation_figures.py`; `python scripts/autoresearch/discovery/validate_predicted_fmri_fold_stability.py --fold expanded20=figures/remote_prediction_inputs_20260428/hcp_language_expanded20_audio_v5 --fold heldout21=figures/remote_prediction_inputs_20260428/hcp_language_heldout21_audio_v1 --out docs/operations/figures/data/hcp_language_predicted_fmri_fold_stability_rerun_20260428.json --n-permutations 20000 --batch-size 256 --seed 20260428`; `python scripts/autoresearch/discovery/validate_hcp_language_barch2013_group_alignment.py --out docs/operations/figures/data/hcp_language_barch2013_alignment_gate_preflight_20260428.json`; and `python scripts/autoresearch/discovery/make_tribe_predicted_response_surface_figures.py --skip-tom --out-root docs/operations/figures/predicted_response_surface_figures_20260428 --fold-stability-json docs/operations/figures/data/hcp_language_predicted_fmri_fold_stability_rerun_20260428.json`. This sequence regenerates Figures 2–14 from the compact CSV/JSON figure data, local ToM prediction artifacts, copied HCP-language prediction-row inputs, the fold-stability rerun JSON, and the Barch-gate preflight JSON. Figure 1 is regenerated separately with `python scripts/autoresearch/discovery/generate_nanobanana_tribe_schematic.py` followed by `python scripts/autoresearch/discovery/overlay_nanobanana_tribe_schematic_labels.py`, or reused from the existing cleaned PNG if no schematic restyling is intended.

The manuscript-render sequence is: `python scripts/autoresearch/discovery/render_tribe_latex_report.py`, followed by copying the generated Markdown, LaTeX source, and PDF to the `/data` mirror. The primary expected local PDF is `docs/operations/latex/tribe_stimulus_discovery_paper_report_2026-04-28/tribe_stimulus_discovery_paper_report_2026-04-28_latex_template.pdf`. The primary expected mirrored PDF is `/data/brain_researcher/research/discovery/docs/operations/tribe_stimulus_discovery_paper_report_2026-04-28_latex_template.pdf`. The PDF validation checks are `pdfinfo` on the local and mirrored PDFs, and `pdftotext` grep checks for `Central hypothesis and registered predictions`, `The decisive experiment`, `Reproducibility and script inventory`, and `References`.

The execution pack outputs should be interpreted by evidence tier. Figures 2–7 are branch-outcome and stimulus-redesign figures. Figures 8–10 and 14 are neural-validation boundary and target-planning figures. Figure 11 is the TRIBE-layer representation result. Figures 12–13 are predicted-response surface diagnostics. The §5.1 preflight JSON records that the TRIBE prediction side is ready but the Barch-2013 group-map gate is blocked by missing external group-map and ROI-mask assets. None of these outputs substitutes for the missing observed subject/run-fold HCP LANGUAGE fMRI target package required by §5.

Manifest and materialization scripts

The manifest and materialization path is recoverable from these repository scripts: `scripts/autoresearch/discovery/manifest_synthesizer.py`; `scripts/autoresearch/discovery/materialize_hcp_ready_runtime.py`; `scripts/autoresearch/discovery/materialize_biomo_runtime.py`; `scripts/autoresearch/discovery/run_biological_motion_redesign_branch.py`; `scripts/autoresearch/discovery/materialize_patch.py`; `scripts/autoresearch/discovery/build_hcp_language_heldout_manifest.py`; `scripts/autoresearch/discovery/build_hcp_language_covariate_sidecar.py`; `scripts/autoresearch/discovery/build_hcp_language_exchangeability_manifest.py`; and `scripts/autoresearch/discovery/build_hcp_language_layer_family_confirmatory_manifest.py`. These scripts define or repair the stimulus manifests, held-out item splits, covariate sidecars, exchangeability strata, and locked layer-family confirmatory manifests used by the reported validations. The biological-motion redesign used `python scripts/autoresearch/discovery/materialize_biomo_runtime.py --walker-mat /data/brain_researcher/research/discovery/inputs/public_protocols/BiologicalMotion/protocol/walkerdata.mat --output-root docs/operations/biological_motion_redesign_20260428/all_walkers_materialized --manifest-path docs/operations/biological_motion_redesign_20260428/biological_motion_dynamic_all_walkers_manifest.json --walker-indices all --duration-seconds 4 --fps 24 --frame-width 512 --frame-height 512 --seed 20260428` followed by `PYTHONPATH=tmp/pydeps_transformers_latest python scripts/autoresearch/discovery/run_biological_motion_redesign_branch.py --manifest docs/operations/biological_motion_redesign_20260428/biological_motion_dynamic_all_walkers_manifest.json --out-root docs/operations/biological_`

motion_redesign_20260428/tribe_predictions_all_walkers --cache-folder tmp/tribev2_cache --device auto --seed 20260428 --max-exact-combinations 50000. The explicit motion-aware representation rerun used MPLCONFIGDIR=/tmp/mplconfig python scripts/autoresearch/discovery/run_biomo_motion_aware_redesign.py; its primary outputs are docs/operations/biological_motion_redesign_20260428/motion_aware_representation/biological_motion_motion_aware_score.json, docs/operations/biological_motion_redesign_20260428/motion_aware_representation/per_item_motion_features.csv, and docs/operations/biological_motion_redesign_20260428/motion_aware_representation/biological_motion_motion_aware_scorecard.png.

Feature extraction and validation scripts

The layer-feature sidecars were generated with scripts/autoresearch/discovery/extract_tribe_layer_features.py. The item- level permutation validators were scripts/autoresearch/discovery/validate_embedding_permutation.py, scripts/autoresearch/discovery/validate_embedding_contrast_permutation.py, and scripts/autoresearch/discovery/validate_embedding_contrast_covariate_adjusted.py. The layer validators were scripts/autoresearch/discovery/validate_layer_feature_contrast_permutation.py and scripts/autoresearch/discovery/validate_layer_feature_family_confirmatory.py. The predicted-response map bridge was checked with scripts/autoresearch/discovery/validate_predicted_fmri_fold_stability.py. The §5.1 cheap intermediate gate was originally bootstrapped by scripts/autoresearch/discovery/validate_hcp_language_barch2013_group_alignment.py (preflight that documented the missing-asset state on 2026-04-28); the executed alignment scripts that produced the verdicts in §5.1 itself are recorded next to their verdict JSONs at figures/data/hcp_language_barch2013_alignment_gate_verdict_20260429.json (Destrieux atlas) and figures/data/hcp_language_barch2013_alignment_gate_hcpmmp_v2_verdict_20260429.json (predeclared HCP-MMP1.0 atlas). Together these scripts separate the exploratory branch scores from the permutation-supported, covariate-adjusted, layer-family, and fold- stability evidence tiers reported above.

Controller and loop guardrail scripts

The operational loop changes that make the branch decisions interpretable are captured by scripts/autoresearch/discovery/proposal_promoter_patch.py, scripts/autoresearch/discovery/score_smoothing_patch.py, scripts/autoresearch/discovery/state_evolution_patch.py, scripts/autoresearch/discovery/oscillation_convergence_patch.py, scripts/autoresearch/discovery/zero_score_refute_patch.py, scripts/autoresearch/discovery/run_action_executor.sh, and scripts/autoresearch/discovery/run_live_watchdog.sh. These are not independent neuroscience evidence. They document the controller rules that prevented no-op follow-ups, repeated seed clones, missing terminal actions, and score-oscillation artifacts from being mistaken for scientific progress, and are the empirical basis of H2.

Figure and report generation scripts

The figure-producing scripts are `scripts/autoresearch/discovery/generate_nanobanana_tribe_schematic.py`, `scripts/autoresearch/discovery/overlay_nanobanana_tribe_schematic_labels.py`, `scripts/autoresearch/discovery/make_tribe_branch_outcome_landscape.py`, `scripts/autoresearch/discovery/make_tribe_remaining_figures.py`, `scripts/autoresearch/discovery/make_tribe_neural_bridge_figures.py`, `scripts/autoresearch/discovery/make_tribe_representation_figures.py`, `scripts/autoresearch/discovery/make_tribe_predicted_response_surface_figures.py`, `scripts/autoresearch/discovery/make_tribe_discovery_story_plate.py`, and `scripts/autoresearch/discovery/make_tribe_discovery_paper_plate_v2.py`. The report was rendered with `scripts/autoresearch/discovery/render_tribe_latex_report.py`, which uses the Brain Researcher LaTeX assets `src/brain_researcher/assets/latex/report_template.tex.j2` and `src/brain_researcher/assets/latex/scientific_report.sty`.

Figure data and image artifacts

The compact figure data tables are `figures/data/branch_outcomes.csv`, `figures/data/hcp_language_evidence.csv`, `figures/data/layer_scores.csv`, and `figures/data/manifest_deltas.json`. The remote prediction-row copies used for figure reproducibility are `figures/remote_prediction_inputs_20260428/hcp_language_expanded20_audio_v5/embedding_rows.jsonl` and `figures/remote_prediction_inputs_20260428/hcp_language_heldout21_audio_v1/embedding_rows.jsonl`.

The final image roots are `figures/nanobanana_schematic_textfree_20260427`, `figures/figure02_branch_outcomes_20260428`, `figures/remaining_figures_20260428`, `figures/neural_bridge_figures_20260428`, `figures/representation_figures_20260428`, and `figures/predicted_response_surface_figures_20260428`.

Report outputs

The editable report is `/home/zijiaochen/projects/brain_researcher/docs/operations/tribe_stimulus_discovery_paper_report_2026-04-28.md`. The mirrored data-directory copy is `/data/brain_researcher/research/discovery/docs/operations/tribe_stimulus_discovery_paper_report_2026-04-28.md`. The Brain Researcher LaTeX-template source is `/home/zijiaochen/projects/brain_researcher/docs/operations/latex/tribe_stimulus_discovery_paper_report_2026-04-28/tribe_stimulus_discovery_paper_report_2026-04-28.tex`. The local LaTeX-template PDF is `/home/zijiaochen/projects/brain_researcher/docs/operations/latex/tribe_stimulus_discovery_paper_report_2026-04-28/tribe_stimulus_discovery_paper_report_2026-04-28_latex_template.pdf`. The mirrored data-directory PDF is `/data/brain_researcher/research/discovery/docs/operations/tribe_stimulus_discovery_paper_report_2026-04-28_latex_template.pdf`.

Figure Legends

Figure 1. Conceptual schematic

This schematic shows the scientific setup and autonomous discovery loop: stimuli, TRIBE model representations or layers, predicted response evidence, contrast scoring, and branch decisions. It is conceptual only and is not a data figure. The image was generated with the Nano Banana text-free schematic generator from a structured prompt; no data are encoded in the figure. AI-generation is disclosed here per the report's limitations (§7).

Image: `figures/nanobanana_schematic_textfree_20260427/tribe_nanobanana_conceptual_schematic_1_labelclean.png`

Figure 2. Branch outcome landscape

Self-driven experiments sort stimulus contrasts into hypothesis classes. The autonomous loop sorted branches into model-tier positive axes, candidate/noisy signals, packaging-sensitive failures, and representation-sensitive redesign candidates. The contrast score is an automated separability measure, not a p-value or direct subject-level neural effect size.

Image: `figures/figure02_branch_outcomes_20260428/figure02_branch_outcome_landscape_20260428.png`

Figure 3. Branch trajectories

Branch decisions are trajectory-dependent. HCP language freezes, HCP social weakens after follow-up and is killed, and auditory remains noisy rather than cleanly reportable.

Image: `figures/remaining_figures_20260428/figure03_branch_trajectories_20260428/figure03_branch_trajectories_20260428.png`

Figure 4. Score decomposition

HCP social did not fail because centroid distance was zero; it failed because directional separation was tiny after a valid follow-up. Only branch states with audited score components are plotted.

Image: `figures/remaining_figures_20260428/figure04_score_decomposition_20260428/figure04_score_decomposition_20260428.png`

Figure 5. Condition signature changes

Real scientific follow-ups require manifest-level condition changes. Math, auditory, ToM, and HCP social illustrate corrected or valid deltas.

Image: figures/remaining_figures_20260428/figure05_condition_signatures_20260428/figure05_condition_signatures_20260428.png

Figure 6. Finding-hypothesis-next-experiment matrix

The discovery loop produces bounded hypothesis classes and concrete next experiments, not only raw branch scores. Hypotheses in this matrix remain bounded by validation status.

Image: figures/remaining_figures_20260428/figure06_finding_hypothesis_matrix_20260428/figure06_finding_hypothesis_matrix_20260428.png

Figure 7. Stimulus redesign roadmap

Positive, noisy, and packaging-sensitive branches imply different next experiments rather than a single rerun policy. The roadmap is a plan, not evidence that future experiments have already run.

Image: figures/remaining_figures_20260428/figure07_redesign_roadmap_20260428/figure07_redesign_roadmap_20260428.png

Figure 8. Neural validation ladder

HCP language has strong item-level and model-feature evidence and late-layer support, but the predicted-fMRI fold bridge is not stable and observed subject-level fMRI validation is missing. This is a validation- tier figure, not an observed fMRI activation map.

Image: figures/neural_bridge_figures_20260428/figure08_neural_validation_ladder_20260428/figure08_neural_validation_ladder_20260428.png

Figure 9. ROI target map

Branch outcomes imply ROI systems to test next, especially auditory and language targets for HCP language and pSTS / TPJ / mPFC / motion targets for social and biological-motion redesigns. Dots indicate future validation targets, not observed activation or ROI-level statistical significance.

Image: figures/neural_bridge_figures_20260428/figure09_roi_target_map_20260428/figure09_roi_target_map_20260428.png

Figure 10. Hypothesis neural-status matrix

HCP language is supported at model-feature, predicted-response, and layer-family tiers, while subject-level observed fMRI and most branch- level ROI claims remain missing or not done.

Image: figures/neural_bridge_figures_20260428/figure10_hypothesis_neural_status_20260428/figure10_hypothesis_neural_status_20260428.png

Figure 11. HCP language layer-family non-replication on n=81 (tier 6b)

Replaces the original Figure 11 (representation profile from the parent campaign, which quoted `T_late_minus_early = +0.859, p ~ 5e-5` on n = 10 selected items). Re-extracting per-layer features locally on the full 81-item validated set and re-applying the **identical predeclared statistic** returns `T = -0.187` (unadjusted) and `T = -3.81` (extended-acoustic adjusted). The original tier-6 result was a small-sample artifact specific to the parent branch's 10-item selection; the late-vs-early ordering is not a stable property of the TRIBE encoder on this stimulus family.

Image: `figures/h1_to_h1prime_figures_20260501/figure11_layer_family_non_replication_20260501.png`

Figure 11b. Per-layer brain alignment with Barch-2013 group activation map

For each TRIBE feature family (16 encoder attention layers, audio projector, text projector), a ridge map fit on 81 items projects the per-layer story-minus-math contrast onto `fsaverage5`; the bar height is the Pearson `r` of that projected contrast against the Barch-2013 group activation map (50 paired NeuroVault subjects averaged, 20,484 vertices). Late encoder attention modules anti-align (mean r = -0.283); early encoder attention is mildly positive (mean r = +0.192); the audio projector at r = +0.644 is the single strongest brain-aligned feature anywhere in TRIBE for this contrast. This figure identifies H1' (audio projector encodes the brain-aligned story-vs-math axis) and dismisses H1 (the original late-encoder narrative-semantic claim) at the brain-alignment level. plus-one permutation `p = 0.0002, N = 5,000`.

Image: `figures/h1_to_h1prime_figures_20260501/figure11b_per_layer_brain_alignment_20260501.png`

Figure 12. IBC ToM predicted fsaverage5 response

Local IBC ToM prediction artifacts can be rendered as TRIBE-predicted `fsaverage5` response contrasts. This is a model-predicted diagnostic only, not observed fMRI or group activation evidence.

Image: `figures/predicted_response_surface_figures_20260428/figure12_tom_predicted_fsaverage5_response_20260428/figure12_tom_predicted_fsaverage5_response_20260428.png`

Figure 13. HCP language predicted fsaverage5 bridge

The two HCP language predicted-response folds can be rendered as `fsaverage5` maps, but their story-minus-math contrast maps are not stable across folds. The map-level bridge is negative (`r = 0.0976`, `Delta r = +0.0936`, `p = 0.449`), and the figure is not observed subject-level fMRI activation evidence.

Image: `figures/predicted_response_surface_figures_20260428/figure13_hcp_language_predicted_fsaverage5_bridge_20260428/figure13_hcp_language_predicted_fsaverage5_bridge_20260428.png`

Figure 14. HCP language support boundary

The HCP language worked example is supported at item-level and layer-family tiers, rejected at the predicted-response fold bridge, blocked at the Barch-2013 group-map gate, and missing observed subject-level fMRI validation. **Note: this figure is the original (pre-rewrite) support-boundary diagram and reflects the campaign state before tier 6b non-replication, tier 6c per-layer brain alignment, and tier 6d subject-level confirmation. It now under-represents what the campaign concluded; see Figures 11, 11b, and 14b for the current state.**

Image: `figures/neural_bridge_figures_20260428/figure14_hcp_language_support_boundary_20260428/figure14_hcp_language_support_boundary_20260428.png`

Figure 14b. H1' rescue at the per-subject level

Per-subject Pearson `r` between TRIBE-feature contrasts and 50 paired HCP S1200 individual story-minus-math zstats (NeuroVault collection 4337). Panel (a) shows audio-projector r (x) vs late-encoder mean r (y) per subject; the green-shaded ideal H1' quadrant (audio-positive / late-negative) contains 49 of 50 subjects. Panel (b) shows per-feature distributions: late encoder mean r = -0.211 (1/50 positive), text projector mean r = -0.043 (14/50 positive), TRIBE final predicted-response output mean r = +0.337 (50/50 positive), audio projector mean r = +0.422 (50/50 positive). The audio projector beats every other feature at the per-subject level, including TRIBE's own published-style downstream output (paired t = +13.7). H1' replicates at every individual subject; H1 (late encoder) is refuted at every individual subject.

Image: `figures/h1_to_h1prime_figures_20260501/figure14b_per_subject_audio_vs_late_20260501.png`

Provenance

The full chronological provenance ledger remains `/data/brain_researcher/research/discovery/docs/operations/br_autoresearch_discovery_complete_unified_2026-04-25.md`. The main validation plan, runtime audit, and layer-feature feasibility notes remain `/data/brain_researcher/research/discovery/docs/operations/hypothesis_discovery_experiment_plan_2026-04-26.md`. The locked follow-up manifest specifications remain `/data/brain_researcher/research/discovery/docs/operations/locked_followup_manifest_specs_2026-04-26.md`. The final figure directory is `/data/brain_researcher/research/discovery/docs/operations/figures`. The local plotting, validation, materialization, and report-rendering scripts live under `/home/zijiaochen/projects/brain_researcher/scripts/autoresearch/discovery`.

The HCP subject-level observed fMRI targets are not included in the current artifact set. This report therefore provides enough provenance to reproduce the item-level model-feature, predicted-response, and figure- generation claims, but it does not provide the observed subject/run-fold neural-validation inputs required by §5. That missing target package remains the required input for the next decisive experiment.

References

d'Ascoli, S., Rapin, J., Benchetrit, Y., Banville, H., and King, J.-R. (2025). TRIBE: TRImodal Brain Encoder for whole-brain fMRI response prediction. arXiv:2507.22229. https://doi.org/10.48550/arXiv.2507.22229

Meta AI. (2026). `facebook/tribev2` model card. Hugging Face. The model card documents TRIBE v2 as a multimodal brain-encoding model for video, audio, and text and describes predictions on the `fsaverage5` cortical mesh. https://huggingface.co/facebook/tribev2

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., Feldt, C., Nolan, D., Bryant, E., Hartley, T., Footer, O., Bjork, J. M., Poldrack, R., Smith, S., Johansen-Berg, H., Snyder, A. Z., Van Essen, D. C., and the WU-Minn HCP Consortium. (2013). Function in the human connectome: Task-fMRI and individual differences in behavior. NeuroImage, 80, 169–189. https://doi.org/10.1016/j.neuroimage.2013.05.033

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., Ugurbil, K., and the WU-Minn HCP Consortium. (2013). The WU-Minn Human Connectome Project: An overview. NeuroImage, 80, 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., and Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. Nature, 536, 171–178. https://doi.org/10.1038/nature18933

Human Connectome Project. (2017). S1200 Group Average Data Release. https://www.humanconnectome.org/study/hcp-young-adult/article/s1200-group-average-data-release

BALSA. (2017). Human HCP 1200 Group Average + Individuals; Structural + fMRI Atlas. https://balsa.wustl.edu/reference/show/pkXDZ

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, 532, 453–458. https://doi.org/10.1038/nature17637

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. Proceedings of the National Academy of Sciences, 118(45), e2105646118. https://doi.org/10.1073/pnas.2105646118

Caucheteux, C., and King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. Communications Biology, 5, 134. https://doi.org/10.1038/s42003-022-03036-1

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. (2024). The AI Scientist: Towards fully automated open-ended scientific discovery. arXiv:2408.06292. https://doi.org/10.48550/arXiv.2408.06292

Romera-Paredes, B., Barekatain, M., Novikov, A., Balog, M., Kumar, M. P., Dupont, E., Ruiz, F. J. R., Ellenberg, J. S., Wang, P., Fawzi, O., Kohli, P., and Fawzi, A. (2024). Mathematical discoveries from program search with large language models. Nature, 625, 468–475. https://doi.org/10.1038/s41586-023-06924-6

Fischl, B. (2012). FreeSurfer. NeuroImage, 62(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., and Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. Frontiers in Neuroinformatics, 8, 14. https://doi.org/10.3389/fninf.2014.00014

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Rio, J. F., Wiebe, M., Peterson, P., Gerard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2