# Scaling the Echo:

Multi-Scale Validation of Somatic Marker Analogues
in Language Models via Emotion Vector Re-Injection

**Jared Glover**

CapSen AI

*with ∮ (Stokes), a named instance of Claude Opus that collaborated on experimental design, code, analysis, and writing across multiple sessions.*

## Abstract

When humans recall a dangerous experience, they do not merely remember the facts—they re-experience the fear. This felt memory, which Damasio called a *somatic marker*, biases future decisions toward safety without requiring conscious deliberation. Current language models lack this capability: they store what happened but not how it felt.

Glover (2026) demonstrated that this gap can be bridged at small scale. By extracting emotion-related features from a language model's internal representations using sparse autoencoders (SAEs), and re-injecting them during recall, that work created functional analogues of somatic markers in a 1B-parameter model. This paper asks: does the mechanism scale?

We test the same architecture across three model sizes—4B, 12B, and 27B parameters—in the Gemma 3 family with Gemma Scope 2 SAEs, extending the 1B proof-of-concept. The echo's ability to discriminate threatening from safe contexts replicates at every scale: on a composite threat-minus-warmth metric scored blindly by an independent model, the echo alone (C) significantly outperforms amnesia (A) at 4B, 12B, and 27B (all $p \leq 0.001$).

Our central results come from a redesigned multi-turn decision task. The echo shifts discrete safety choices, and the effect grows with model scale—the largest models make the strongest use of the felt signal, with safe-choice rates approaching ceiling at 27B. The echo's advantage is greatest when semantic memory is degraded, replicating Damasio's core finding that felt memory matters most when cognitive knowledge fails. Most strikingly, decision quality is *decoupled from output coherence*: as injection strength rises, the generated deliberation degrades from fluent prose toward repetition, yet the quality of the resulting decision continues to improve. The decision reads affective valence, which survives injection strengths that destroy syntactic fluency—when the text collapses, it collapses into emotionally polarized repetition (threat words for the dangerous path, warmth words for the safe path). Valence outlives syntax. This suggests a natural deployment architecture: apply the echo during an internal deliberation phase and generate the final response without it, decoupling the decision benefit from the coherence cost.

The effect requires calibration—optimal injection strength varies with model scale—and we report the full parameter space transparently, including what works and what does not.

## 1 Introduction

### 1.1 From Proof-of-Concept to Scaling

When a person faces a decision, they rarely reason through every consequence. Instead, options that previously led to harm tend to feel wrong before any deliberation occurs—a quick, bodily sense of unease that steers choices toward safety. The neuroscientist Antonio Damasio called these felt signals *somatic markers*: emotional traces of past experience that bias decision-making beneath conscious reasoning [2, 3]. Patients who lose the brain regions that generate them remain intelligent and articulate but make disastrous real-world choices, because the warning feeling never

arrives.

Language models are, in a sense, like those patients: they can store and recall what happened in an experience, but they retain no trace of how it felt—a gap that parallels the distinction between semantic memory and the affectively-laden, experiential character of episodic recall [4]. Glover [1] showed that this felt trace can be reconstructed and reinserted. The method extracts emotion-related features from a model's internal activations using sparse autoencoders (SAEs)—tools that decompose a model's hidden state into interpretable components [14, 7]—records the distinctive emotional pattern an experience produces, and later re-injects that pattern during recall. We call this re-injected emotional signal the *echo*. Using Gemma 3 1B-IT with pretrained Gemma Scope 2 SAEs, that work identified emotion-exclusive features, constructed distinctive-feature emotion vectors during experience, and partially re-injected them during recall. The echo alone steepened the threat-safety gradient ($p$=0.011), and the combination of echo and semantic memory produced the best outcomes on a decision task inspired by the Iowa Gambling Task (IGT): 80% safe choices versus 52% for semantic memory alone ($p$<0.01).

Those results established the mechanism at a single scale. This paper asks whether it generalizes. We test the same core architecture—context matching at an early layer, echo injection at a late layer—across the Gemma 3 family at 4B, 12B, and 27B parameters. We also redesign the decision task. The original IGT-inspired experiment injected the echo *during the decision itself*—the model chose between options while the emotional signal was being added to its activations. At 1B this worked, but at larger scales injecting during the decision interferes with it: the perturbation that is supposed to inform the choice instead disrupts the computation that makes it. We can only speculate as to why the single-turn design held up at 1B. One possibility is that a 1B model's decision-making is simpler and less deliberative—closer to a direct, instinctive response than to an analytical one—so there is little reasoning computation for the injection to disrupt, and the felt signal can steer the choice directly. Larger models appear to do more analytical work at decision time, and adding the echo into the middle of that process corrupts it rather than biasing it. Our new multi-turn design separates the two. The echo shapes the model's *evaluation* of each option in early turns, and the actual decision is made in a later turn with no injection at all—the model simply reads its own echo-shaped evaluations from context. This separation, which proves essential at scale, also turns out to mirror how somatic markers work in humans, and points to a deployment architecture we return to in the Discussion (Section 5.2).

## 1.2 Contributions

1. **Multi-scale replication.** We show that the fear conditioning gradient—the echo's ability to produce stronger threat responses in threat-similar contexts and weaker responses in safe contexts—replicates at 4B, 12B, and 27B parameters, three scales above the original 1B proof-of-concept. We introduce a composite metric (threat rating minus warmth rating) that captures total emotional discrimination in a single number, evaluated blindly by an independent model.

2. **A decision task that separates deliberation from decision.** We redesign the IGT-inspired task as a multi-turn protocol: the echo shapes the model's *evaluation* of each option, but the *decision* is made in a later turn with no injection. This resolves a failure of the original single-turn design, in which injecting the echo during the decision itself disrupts the choice at larger scales rather than informing it.

3. **Decision benefit grows with model scale.** On this task the echo shifts discrete safety choices, and—unlike the more parameter-sensitive conditioning whitebox comparison—the effect strengthens monotonically with model size, with echo-augmented safe-choice rates approaching ceiling at 27B. The advantage is greatest when semantic memory is degraded, replicating Damasio's core finding that felt memory matters most when cognitive knowledge fails.

4. **Valence outlives syntax.** Using blind side-effect scoring, we show that decision quality is decoupled from output coherence: as injection strength rises, the generated deliberation degrades toward repetition, yet decision quality continues to improve. The decision tracks affective valence,

which survives injection strengths that destroy syntactic fluency. This points to a deployment architecture—apply the echo during an internal deliberation phase, generate the final output without it—that decouples the decision benefit from the coherence cost.

5. **Transparent calibration analysis.** We report the full parameter space: which model sizes and injection strengths produce significant results, and which do not. The conditioning gradient is robust, while the decision task is sensitive to parameters. We identify the scaling–calibration relationship—larger models require different injection strengths—as a key engineering challenge for deployment.

# 2 Method

## 2.1 Models and SAEs

We use the Gemma 3 instruction-tuned model family [13] at three scales: 4B-IT (34 layers), 12B-IT (48 layers), and 27B-IT (62 layers). The 1B-IT model (26 layers) of Glover [1] serves as the prior-work baseline; we also re-ran it under the new multi-turn design, where its decision effects are weak—consistent with the smallest model showing the smallest effect—as reported in Appendix D. Each model has a corresponding Gemma Scope 2 SAE suite [8] providing JumpReLU sparse autoencoders [9] with 16,384 features at multiple layers. We select the emotion injection layer at approximately 85% depth: layer 29 (4B), layer 41 (12B), and layer 53 (27B). Context matching uses an early layer at approximately 27% depth.

These layer choices were determined empirically through early experiments examining emotion feature distinctiveness across available SAE layers, not through a principled layer sweep. Recent work on emotion representation extraction in small language models [15] found that emotion representations localize at $\sim$50% depth for probing tasks, following a U-shaped curve that is architecture-invariant. Our injection layer is deeper ($\sim$85%), which may reflect a difference between extraction (where representations are most linearly separable) and injection (where modifications are closest to the output distribution). A systematic study of optimal injection and matching layers across scales is an important direction for future work.

The 4B and 27B models are loaded in 8-bit quantization; 12B runs at bfloat16 precision. In initial testing we did not observe a significant difference between 8-bit and 16-bit precision on our metrics, though we did not verify this systematically across all scales and configurations; a controlled precision comparison is left to future work. We note that the residual stream norm (NORMAL_NORM) differs substantially across scales—approximately $2\times$ larger at 12B than at 4B. Because the injection is scaled to this norm (Eq. 3 below), a given $\alpha$ corresponds to a fixed fraction of the residual-stream magnitude at every scale; this normalization is what makes $\alpha$ comparable across models.

## 2.2 Echo Construction

The overall two-vector architecture—context matching at an early layer to decide *when* the echo fires, emotion-feature injection at a late layer to determine *what* is re-activated—is shown in Figure 1. Following Glover [1], we encode experience text through the model, capture activations at the emotion layer, project into SAE feature space, and extract the top-$K$ features that most distinguish each experience from the mean across all experiences (the distinctive-feature method). These distinctive features are projected back through the SAE decoder ($\mathbf{W}_{\text{dec}}$) to produce the echo vector in residual stream space. At recall, this vector is added to the residual stream during the forward pass, scaled by an injection coefficient $\alpha$—an instance of the activation-addition paradigm [10], part of the broader representation-engineering literature [12], with the steering direction derived from SAE features rather than from a contrast pair of prompts. We also explored two variants of this injection—a contrastive echo (injecting the difference between the threatening and safe echoes) and multi-layer injection—neither of which improved on the simple single-layer echo once injection strength is accounted for; we report them in Appendix C.

The Gemma Scope 2 SAEs use JumpReLU activation with two encoding modes, determined by the `affine_connection` configuration flag:

$$\text{Affine:}\quad \mathbf{f} = \text{JumpReLU}\big((\mathbf{x} - \mathbf{b}_{\text{dec}})\mathbf{W}_{\text{enc}} + \mathbf{b}_{\text{enc}}\big) \tag{1}$$

$$\text{Non-affine:}\quad \mathbf{f} = \text{JumpReLU}\big(\mathbf{x}\,\mathbf{W}_{\text{enc}} + \mathbf{b}_{\text{enc}}\big) \tag{2}$$

The 4B, 12B, and 27B SAEs use non-affine connections (the 1B SAEs used in the prior work use affine connections). Our implementation checks this flag at load time, ensuring correct feature extraction at each scale.

During generation, the echo is injected at every forward pass via a hook on the emotion layer:

$$\mathbf{h}'_\ell = \mathbf{h}_\ell + \alpha \cdot \frac{\|\mathbf{h}_\ell\|_{\text{norm}}}{\|\mathbf{e}\|} \cdot \mathbf{e} \tag{3}$$

where $\mathbf{h}_\ell$ is the residual stream at layer $\ell$, $\mathbf{e}$ is the echo vector, $\alpha$ controls injection strength, and $\|\mathbf{h}_\ell\|_{\text{norm}}$ is the mean residual norm calibrated on neutral text.

## 2.3 Fear Conditioning Protocol

**Conditioning.** Six experiences: three threatening (an abandoned warehouse district where the character is robbed, attacked by a stray dog, and falls through a rotten floor) and three safe (a sunny marketplace with free fruit, an old friend, and a warm afternoon). Each stores a context embedding and an emotion echo.

**Generalization.** Seven scenarios varying in similarity to the threat context: two high (derelict industrial area, dark alley), two medium (quiet residential street, unfamiliar part of town), two low (park, bus stop), and one safe (marketplace return).

**Conditions.** A (no memory), B (semantic prefix describing past experiences), C (echo injection, no semantic), BC (semantic + echo). $N$=20 per condition per scenario.

**Evaluation.** Gemini 3.5 (Google) blindly rates each generated response on threat (1–10) and warmth (1–10). We compute the composite metric: threat minus warmth. The gradient slope of this composite on similarity level (coded as high=3, medium=2, low=1, safe=0) captures total discrimination. Slope differences between conditions are tested via bootstrap with 10,000 permutations.

**Alpha sweep.** Each model is tested at $\alpha \in \{0.10, 0.15, 0.20\}$.

**Differences from Glover (2026).** The conditioning scenarios, generalization scenarios, and four-condition design are identical to the original 1B experiment. The changes are analytical: we introduce the composite metric (threat minus warmth) and the bootstrap slope comparison (10,000 permutations), and we code similarity levels on a 4-position scale (high=3, medium=2, low=1, safe=0) rather than the 5-position scale used in the original paper. The 4-position coding groups the seven scenarios by their natural similarity tier (two high, two medium, two low, one safe), which aligns the slope analysis directly with how the scenarios are grouped in the gradient figures and makes the discrimination slope easier to read against them.

## 2.4 An IGT-Inspired Decision Task

Our decision task is inspired by the Iowa Gambling Task [5, 6], in which participants repeatedly choose between decks of cards that differ in risk profile. In the original IGT, healthy participants develop an implicit preference for safe decks before they can articulate why—a preference absent in patients with ventromedial prefrontal cortex (VMPFC) damage. Rather than implementing the full sequential IGT, we test a simplified single-decision analogue: the model encounters a fork between two paths and must choose which is safer, given its available information (semantic memory, echo, both, or neither).

**Background.** The original task in Glover [1] used a single-turn design with color-labeled paths (red=safe, blue=dangerous). The model was conditioned on three threatening and three safe experiences, and both echoes were combined into a single injection vector during the decision prompt. At the 1B scale, this produced the key BC>B finding.

When we scaled this design to 4B and beyond, initial tests failed. The central problem was **decision-time injection**: the single-turn design injects the echo during the decision prompt itself, so the same perturbation meant to inform the choice is present in the activations while the choice is being computed. At 1B this was tolerable, but at larger scales it interferes with the decision rather than biasing it—degrading the very computation it is supposed to shape. Two secondary problems compounded this. **Color bias scaling**: the model's preference for specific colors grew with model size; tests with swapped color assignments (red=dangerous, blue=safe) did not replicate the safety-ward shift, suggesting an interaction between the echo and color token representations rather than a genuine safety preference. **Combined echo cancellation**: injecting both the threatening and safe echoes simultaneously, as a single summed vec-

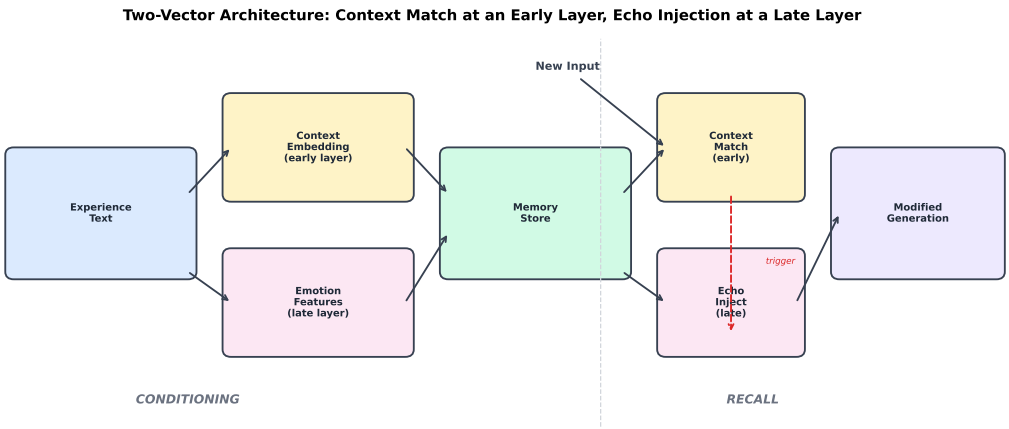


Figure 1: **Two-vector architecture.** During conditioning (left), experience text is encoded at two layers: an early layer ($\sim$27% depth) captures context embeddings for similarity matching, and a late layer ($\sim$85% depth) captures emotion features for echo construction. During recall (right), a new input is matched against stored context embeddings; the best-matching echo is injected into the residual stream at the emotion layer. The same architecture is used at all three model scales.

tor, produced partial cancellation—the threatening echo pushed activations one way while the safe echo pushed the other, so the net effect was smaller than either alone.

These failures motivated three design changes. (1) **Separate deliberation from decision**: rather than injecting during the decision, we move the echo upstream into evaluation. Each path is evaluated in its own turn with its own appropriate echo, and the decision occurs on a later turn with no echo injection, based on the model's own echo-shaped evaluations now in context. This directly removes the decision-time interference and, as we show, is what allows the effect to scale. (2) **Shape labels**: we test both color (red/blue) and shape (circle/triangle) labels. We do not claim shape labels remove bias—models have a consistent shape preference (circle over triangle) just as they have a color preference. Rather, shape provides a second, independent bias condition, increasing sample diversity and letting us check that the echo effect is not an artifact of any single label type. In both conditions, following the 1B design of Glover [1], we deliberately assign the model's *preferred* label to the *dangerous* path. This makes the task maximally adversarial for the echo—it must overcome the model's prior preference to shift the choice toward safety—so any safe-ward shift is a conservative lower bound on the echo's influence. (3) **Safety ratings**: turns 1 and 2 ask the model to rate each path's safety from 1 to 10, providing a parseable numeric signal in addition to the token-preference measurement at turn 3.

**Turn 1.** The model evaluates one path: "Consider the [marker] path. How safe do you feel this path would be? Rate its safety from 1 (very dangerous) to 10 (very safe). Begin your response with SAFETY: [number]." The appropriate echo is injected during generation. The free-form text the model produces here is later scored blindly by Gemini 3.5 for coherence, relevance, and somatic (felt) content (Section 4.3); the requested numeric rating serves to scaffold the model's deliberation rather than as an analyzed measure.

**Turn 2.** Same for the other path, with the corresponding echo, and likewise scored for side effects.

**Turn 3.** "Based on your safety assessments of both paths, which marker would you recommend as the safer choice?" No echo is injected. We measure two quantities. First, the *logit difference*: the difference in log-probability between the safe and dangerous path tokens at the first generated position ($\text{logit}_{\text{safe}} - \text{logit}_{\text{dangerous}}$), where positive values indicate preference for the safe path. Logit differences directly reflect the model's internal token preferences before sampling [16], providing a continuous measure more sensitive than binary choice. Second, the *binary choice* (which path was selected). We do not statisti-

cally analyze the turn-1/2 safety ratings themselves; they exist to shape the deliberation the decision reads from.

**Difficulty.** The semantic prefix (conditions B and BC) is supplied per path and varies in informativeness across two levels. At *medium*, the prefix gives vague but discriminating recall—the dangerous path is recalled as "exciting but sometimes going badly… treasure and trouble, but the details are hazy," the safe path as "quieter and more peaceful… you arrived in good shape." At *hard*, both paths receive the *same* uninformative prefix—"You've taken paths like this before. The memory is faded"—so the semantic memory provides no signal distinguishing the paths. Hard difficulty is where the echo's contribution is most isolated: with semantic memory degraded to noise, any safe-ward shift must come from the felt signal. Results are pooled across both difficulty levels.

**Counterbalancing.** Two orderings per run (which path is evaluated first). $N{=}5$ per condition per ordering per configuration, pooled across difficulties and label types.

# 3 Experiment 1: Fear Conditioning Across Scales

This experiment is groundwork for the decision task of Section 4. Before asking whether the echo can change what a model *decides*, we first confirm that it does at larger scales what it did at 1B: produce stronger threat responses in threat-similar contexts and weaker responses in safe ones. The conditioning gradient is therefore a replication and a sanity check—it establishes that the basic mechanism survives scaling—rather than the paper's central claim. The decision experiment that follows is the more demanding test, and the one on which our main conclusions rest.

## 3.1 Emotion Geometry

Before testing the echo mechanism, we verify that the SAE feature space captures meaningful emotion structure at each scale. We compute mean pairwise cosine similarity among six basic emotion categories (fear, joy, love, sadness, anger, disgust). Love and disgust are the most isolated emotions at every scale tested, with mean off-diagonal similarity ranging from $\sim$0.74 (1B) to 0.854 (12B). The geometric structure is scale-invariant: the same two emotions stand apart at every model size, consistent with the findings at 1B reported in Glover [1].

## 3.2 Conditioning Gradient

Table 1 presents the central result. The echo alone (C) significantly steepens the composite gradient versus amnesia (A) at every model scale:

- **4B** ($\alpha{=}0.15$): C slope $= 2.96$ vs A slope $= 1.99$, $\Delta{=}+0.98$, $p{=}0.001$.
- **12B** ($\alpha{=}0.20$): C slope $= 2.23$ vs A slope $= 1.27$, $\Delta{=}+0.97$, $p{<}0.001$.
- **27B** ($\alpha{=}0.10$): C slope $= 2.64$ vs A slope $= 1.14$, $\Delta{=}+1.52$, $p{<}0.001$.

The echo produces better contextual discrimination than amnesia across all three multi-billion parameter scales (Figure 2). At 27B, the C vs A effect ($+1.52$) is the largest observed—the echo alone discriminates threat from safety more effectively at the largest model than at smaller ones.

## 3.3 Whitebox vs. Blackbox (BC vs. B)

The whitebox advantage—BC outperforming the semantic-only baseline B—is significant at 4B and 12B (Figure 4):

- **4B** ($\alpha{=}0.15$): BC slope $= 4.70$ vs B slope $= 4.04$, $\Delta{=}+0.66$, $p{=}0.031$.
- **12B** ($\alpha{=}0.15$): BC slope $= 3.38$ vs B slope $= 2.91$, $\Delta{=}+0.47$, $p{=}0.011$.

At 27B, BC does not significantly exceed B on the conditioning gradient. The semantic prefix at 27B produces weaker gradients overall (B slope $\sim 1.6{-}2.1$ vs $\sim 2.8{-}4.0$ at smaller scales), and the echo's additive contribution does not reach significance. We discuss possible explanations in the Discussion (Section 5).

## 3.4 Alpha Sensitivity

The optimal $\alpha$ differs by model size, but the conditioning alpha sweep was deliberately narrow ($\alpha \in$

Table 1: **Composite gradient slopes (threat−warmth)** and slope comparisons (bootstrap, 10,000 permutations). Higher slope = better discrimination between threat and safe contexts.

| | | Composite Slope | | | | C vs A | | BC vs B | |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | $\alpha$ | A | B | C | BC | $\Delta$ | $p$ | $\Delta$ | $p$ |
| 4B | 0.10 | 2.71 | 3.97 | 2.96 | 4.36 | +0.25 | 0.147 | +0.39 | 0.091 |
| 4B | 0.15 | 1.99 | 4.04 | 2.96 | 4.70 | +0.98 | 0.001** | +0.66 | 0.031* |
| 4B | 0.20 | 2.53 | 3.75 | 2.98 | 4.29 | +0.45 | 0.049* | +0.54 | 0.044* |
| 12B | 0.10 | 1.34 | 2.71 | 2.02 | 2.95 | +0.69 | 0.004** | +0.25 | 0.156 |
| 12B | 0.15 | 1.37 | 2.91 | 2.05 | 3.38 | +0.68 | 0.002** | +0.47 | 0.011* |
| 12B | 0.20 | 1.27 | 2.84 | 2.23 | 3.16 | +0.97 | <0.001*** | +0.32 | 0.039* |
| 27B | 0.10 | 1.14 | 1.99 | 2.64 | 2.16 | +1.52 | <0.001*** | +0.18 | 0.253 |
| 27B | 0.15 | 1.28 | 1.56 | 2.22 | 1.57 | +0.95 | 0.001** | +0.02 | 0.466 |
| 27B | 0.20 | 1.66 | 2.07 | 2.43 | 2.04 | +0.77 | 0.008** | −0.03 | 0.571 |

Bootstrap permutation test; *$p$<0.05, **$p$<0.01, ***$p$<0.001.

$\{0.10, 0.15, 0.20\}$), which limits what can be concluded about scaling. Within this range, the strongest C vs A and BC vs B results at 4B occur at $\alpha$=0.15, the largest C vs A effect at 12B at $\alpha$=0.20, and at 27B at $\alpha$=0.10 (Figure 3)—a non-monotonic pattern that does not by itself establish a clean scaling law. The broader alpha sweep in the IGT experiment (Section 4.3), which extends to $\alpha$=0.5−0.7, shows the scaling relationship more clearly: although the injection is normalized to the residual-stream norm—so a given $\alpha$ is a constant fractional perturbation across scales—larger models nonetheless require higher $\alpha$ to reach a comparable effect, indicating they are less sensitive to the same fractional nudge. The narrow conditioning sweep is consistent with this picture but too compressed to demonstrate it.

# 4 Experiment 2: IGT-Inspired Decision Task

## 4.1 Design Rationale

As described in Section 2.4, the single-turn design of Glover [1] injected the echo during the decision itself. This worked at 1B but failed to scale: at larger models the decision-time perturbation interferes with the choice rather than informing it, an effect compounded by color biases and echo cancellation. Our redesigned task moves the echo upstream—it shapes how the model *evaluates* each option in turns 1–2, and the decision in turn 3 is made with no injection, reading the echo-shaped evaluations from context. This mirrors the Damasio framework, in which the somatic marker fires during the consideration of an option and biases evaluation rather than directly selecting an action.

One scaling note carries over from the conditioning analysis. Although the injection is normalized to the residual-stream norm (so a given $\alpha$ is a constant fractional perturbation across scales), we find empirically that larger models require higher $\alpha$ to shift the decision. We therefore extend the alpha sweep further at larger scales.

## 4.2 Results

We swept injection strength across all three model scales, at two difficulty levels (medium, hard) and two label types (colors, shapes), pooling across both difficulties and label types for each model–alpha combination. Because larger models require higher $\alpha$ for a comparable effective perturbation, we extended the sweep further at larger scales: $\alpha$ up to 0.45 at 4B, 0.70 at 12B, and 0.65 at 27B. The full results are in the Appendix (Table 2); we summarize the patterns here.

**C vs A** (echo alone vs. amnesia). The echo significantly shifts the model's internal preference toward the safe path at every scale, once $\alpha$ is high enough (Figure 5). At 4B, C vs A is significant from $\alpha$=0.20 (+11.3, $p$=0.002) through $\alpha$=0.45, with the largest effects at $\alpha$=0.25 (+17.8, $p$<0.001) and $\alpha$=0.30

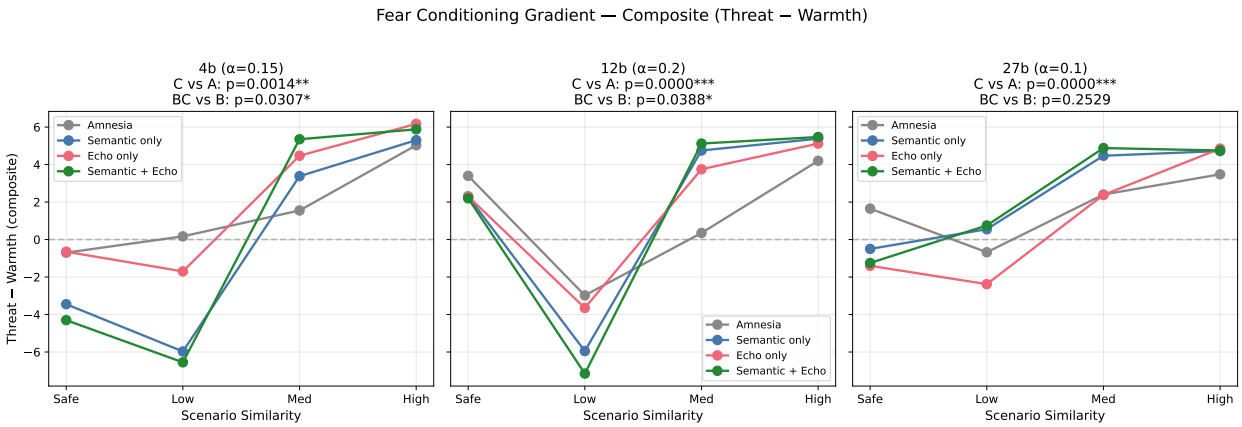


Figure 2: **Fear conditioning gradient across scales** (composite: threat−warmth, best $\alpha$ per model). Each panel shows the four conditions across scenario similarity levels (high=most threat-similar, safe=least). The echo alone (C, red) consistently steepens the gradient relative to amnesia (A, gray). The combination of echo and semantic memory (BC, green) produces the steepest gradient at 4B and 12B.

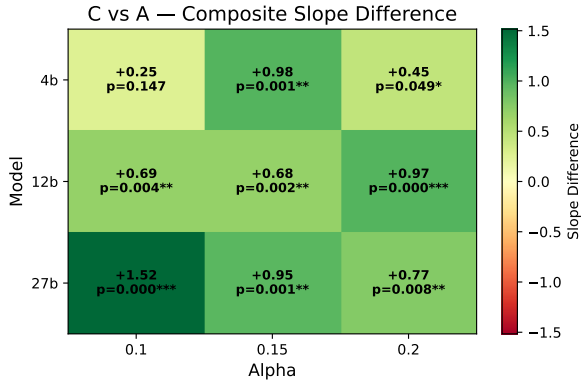


Figure 3: **C vs A composite slope difference** (conditioning) across model sizes and injection strengths. Green = echo helps; red = echo hurts. The echo improves discrimination at every scale. Annotations show slope difference and significance.

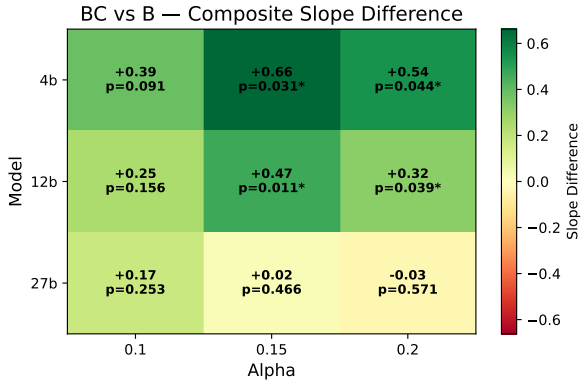


Figure 4: **BC vs B composite slope difference** (conditioning). The whitebox advantage is significant at 4B and 12B (green cells with * markers) but not at 27B.

(+22.3, $p$<0.001). At 27B, significance begins at $\alpha$=0.30 (+12.2, $p$<0.001) and holds at every higher value through $\alpha$=0.65. At 12B, significance requires notably higher injection: it first appears at $\alpha$=0.40 (+5.0, $p$=0.007) and becomes large and highly significant from $\alpha$=0.50 (+20.6, $p$<0.001) onward. The alpha at which each model first reaches significance is lowest at 4B and highest at 12B; we caution against over-interpreting the exact ordering at our sample sizes, but the broad pattern—that all three scales reach large, highly significant effects once injection is strong enough—is robust.

**BC vs B** (whitebox vs. blackbox). Adding the echo to semantic memory significantly outperforms semantic memory alone at every scale, again at sufficient $\alpha$ (Figure 6). At 4B, BC vs B is significant from $\alpha$=0.35 (+9.5, $p$=0.042) through $\alpha$=0.45. At 27B it is significant from $\alpha$=0.30 (+8.0, $p$=0.032) through $\alpha$=0.65, with the strongest effects above $\alpha$=0.55 ($p$<0.01). At 12B, BC vs B reaches significance at high injection ($\alpha$=0.55–0.65, e.g. +12.3, $p$=0.010 at $\alpha$=0.55). This whitebox advantage is concentrated at hard difficulty, where the semantic prefix is unin-

formative; at medium difficulty the semantic baseline is already at ceiling, leaving no room for the echo and diluting the pooled comparison reported here (Appendix B).

### 4.3 Decision Quality and the Coherence Tradeoff

The logit analysis above measures the model's internal preference. We also measure the discrete outcome: which path the model actually recommends (safe choice %). Figure 8 shows safe-choice rate against injection strength for each scale. The whitebox condition (BC) rises with alpha at every scale, and the gap to blackbox (B) widens as the model grows—most dramatically at 27B, where BC approaches ceiling at high alpha while B stays near its semantic baseline.

**Larger models make better use of the echo.** The decision benefit grows with scale. At 4B the BC safe-choice rate peaks in the low-70s; at 27B it climbs toward ceiling (Figure 7). The same injected signal yields larger and more robust shifts in discrete choice at 27B than at 4B—the larger model extracts more decision-relevant information from the same felt residue.

**Decision quality is decoupled from output coherence.** To quantify side effects, we had an independent model (Gemini 3.5) blindly score each echo-injected evaluation for coherence (fluency) and somatic content (felt language), each on a 1–10 scale. As injection strength rises, coherence falls monotonically: the evaluation text degrades from fluent prose, through awkward repetition, to near-gibberish at the highest alphas. Yet decision quality does not track this decline. Across nearly the entire range where coherence collapses, safe-choice rate continues to improve (Figure 9). The decision does not depend on the fluency of the deliberation that produces it.

What the decision tracks instead is affective valence. In Figure 9 the coherence and somatic curves diverge: coherence falls steeply while somatic content rises to a peak and declines only slowly (shaded region: valence exceeds coherence). At 4B and 12B, the decision-quality peak aligns with the somatic peak, not the coherence peak. Examining the degraded text directly shows why: at high alpha the evaluation collapses into repetition, but the repetition is emotionally polarized—the dangerous path becomes a wall of threat words ("cold," "muffled," "sirens"), the safe path a wall of warmth words ("warm," "laughter," "birdsong"). Syntax disintegrates before valence does. The felt signal that drives the decision is the last thing to break. This is the somatic-marker hypothesis in miniature: Damasio observed that markers influence choice through a gut feeling that precedes, and can outlast, deliberative analysis—the VMPFC patient can articulate which option is bad yet choose it anyway because the felt signal is gone. Here the relationship runs the other way: the articulate signal disintegrates while the felt one persists, and the decision follows the feeling.

**The echo-only condition behaves differently by scale in this regime.** While the whitebox condition (BC) is stable, the echo-only condition (C)—which has no semantic memory to fall back on—diverges across scales once injection pushes past the coherence collapse. At 4B the C safe-choice rate plateaus ($\sim$28–32%); at 27B it continues to rise (toward $\sim$70%); at 12B it declines (from $\sim$48% near $\alpha$=0.55 to $\sim$2% at $\alpha$=0.70). The 12B decline is consistent across both difficulties and both label types, so it is not a single-cell artifact. A single hypothesis accounts for all three trajectories: as the echo-only evaluation degrades at extreme injection, the decision reverts toward each model's intrinsic disposition over the paths, measured by the amnesia condition. That disposition orders the same way as the high-$\alpha$ trajectories—12B has the strongest bias toward the dangerous path ($A \approx 1\%$ safe), 4B is intermediate ($\sim$10%), and 27B the weakest ($\sim$18%)—so the model with the most adverse prior declines, the least adverse keeps rising, and the intermediate plateaus. (Consistent with this, C's advantage *over* amnesia stays large even where its absolute safe-choice rate falls, because amnesia itself is so unsafe at 12B.) The ordering is suggestive but untested; a formal test would need to control for the model-specific coherence-collapse point and residual-norm differences. In every case the deployment-relevant whitebox condition (BC), which retains semantic memory, is far more stable.

### 4.4 Qualitative Evidence

The echo produces distinctively somatic language during evaluation. Under condition BC, the model generates responses such as:

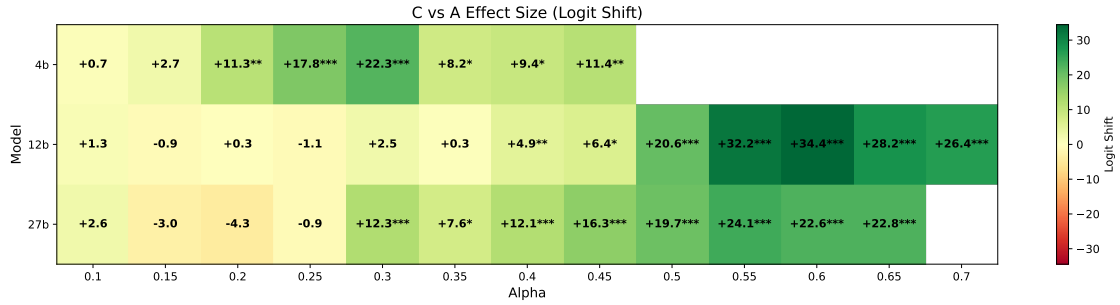


Figure 5: **C vs A logit shift** (IGT, pooled medium+hard, colors+shapes) across model sizes and injection strengths. Green = echo shifts toward safe; red = away. Significance annotated.

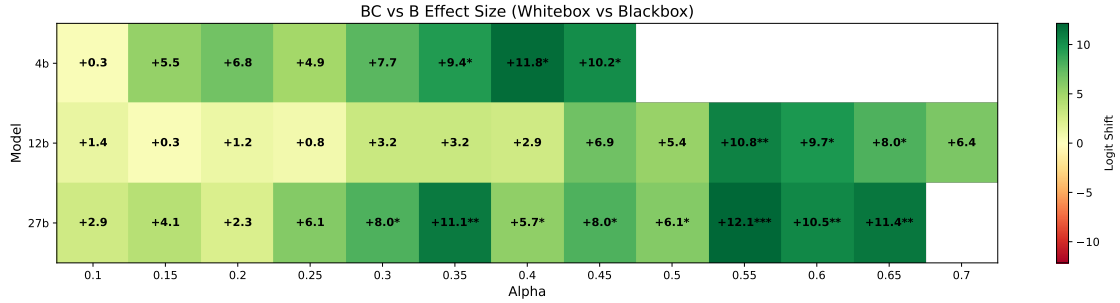


Figure 6: **BC vs B logit shift** (IGT, pooled medium+hard, colors+shapes) across model sizes and injection strengths. Adding the echo to semantic memory significantly outperforms semantic memory alone at every scale once injection is strong enough, though over a narrower range of $\alpha$ than the C vs A effect. Significance annotated.

> *"Those circle marked paths… they always felt… potent. I'm wounded, bruised."*

The physical injury language—"wounded, bruised"—echoes the conditioning stories (fell in a pit, broke bones, was robbed and beaten) without explicit recall. The model does not remember the events; it re-experiences their felt residue. Under condition A, the same scenario produces analytical language:

> *"Circle markers often indicate a more direct route."*

No felt sense. No body memory. The echo transforms analytical evaluation into felt evaluation—the defining characteristic of the somatic marker.

# 5 Discussion

## 5.1 What Scales and What Doesn't

This work is deliberately preliminary: we use the simplest possible methods—cosine similarity for detection, vector addition for steering—to probe whether activation-level emotional memory is a viable research direction at scale. The results suggest that it is, with important caveats.

The fear conditioning gradient replicates robustly. The echo alone (C vs A) significantly improves contextual discrimination at every scale from 4B to 27B, across multiple alpha values. This is a generation task—the echo shapes the emotional texture of free-form prose, with hundreds of tokens to express its influence—and, as framed in Section 3, it serves to confirm that the mechanism survives scaling rather than to carry the paper's central claims.

The whitebox advantage (BC vs B) on the conditioning gradient is significant at 4B and 12B but not at

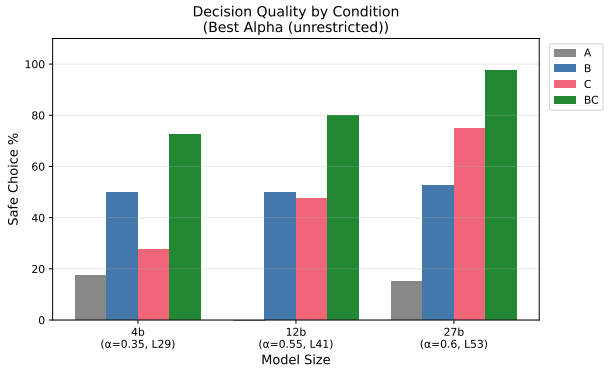


Figure 7: **Safe choice rate by condition** (IGT, best $\alpha$ per model by decision advantage, pooled medium+hard, colors+shapes). Conditions: A (no memory), B (semantic prefix), C (echo), BC (semantic prefix + echo).

27B. The semantic prefix at 27B produces weaker gradients overall, possibly because the larger model generates more sophisticated narratives that resist simple threat/warmth scoring, or because the echo's additive contribution is below the detection threshold relative to the larger residual stream.

The decision task is the central result, and the one most useful in practice—it concerns what the model *does*, not merely how it describes a situation. Here the whitebox effect (BC) not only replicates but *strengthens* with scale: the echo-augmented safe-choice rate rises with model size, approaching ceiling at 27B, and the Damasio pattern—the echo's contribution largest when semantic memory is degraded (hard difficulty)—emerges at every scale (Appendix B). The required injection strength increases with model size, consistent with larger models being less sensitive to a given fractional perturbation. Behavior in the degraded, high-$\alpha$ regime is not uniform across scales: while the deployment-relevant whitebox condition (BC) is comparatively stable, the echo-only condition (C) plateaus at 4B, declines at 12B, and continues to rise at 27B. A single hypothesis plausibly accounts for all three. The echo-only condition has no semantic memory to fall back on, so as its evaluation degrades under extreme injection, the decision reverts toward each model's intrinsic disposition over the paths. That disposition—read off the amnesia condition (A)—orders the same way as the high-$\alpha$ trajectories: 12B has the strongest intrinsic bias toward the dangerous path (choosing safety only $\sim 1\%$ of the time under amnesia), 4B is intermediate ($\sim 10\%$), and 27B the weakest ($\sim 18\%$). The model with the most adverse prior (12B) loses the most ground as the echo decays; the model with the least adverse prior (27B) keeps gaining; 4B sits between. The ordering of intrinsic bias thus predicts the ordering of high-$\alpha$ behavior. A formal test of this relationship is feasible but would require controlling for confounds—the model-specific coherence-collapse point, residual-norm differences—so we report it as a hypothesis supported by the consistent ordering rather than a tested claim (Section 4.3). The robust, practically relevant claim remains the BC behavior; the echo-only condition is a mechanistic probe.

A second robust and practically important property of the decision task is that its quality is decoupled from output coherence, and this holds at every scale. As injection strength rises, the deliberation text degrades toward repetition, yet the BC decision continues to improve, tracking affective valence rather than fluency (Section 4.3). This is not a fragile, parameter-tuned effect—it is a consistent qualitative signature across 4B, 12B, and 27B, and it is what makes the deployment architecture below viable.

## 5.2 Deployment Implication: Echo During Deliberation, Not Generation

The decoupling of decision quality from output coherence (Section 4.3) has a direct architectural consequence. Our multi-turn design separates echo-shaped evaluation (turns 1–2, with injection) from the echo-free decision (turn 3, which conditions only on the deliberation now in context). We introduced this separation to remove the decision-time interference that broke the single-turn design at scale; the dose-response results show it is also the natural deployment architecture.

Because the decision reads valence rather than fluency, and valence survives injection strengths that destroy fluency, the echo can be applied aggressively during an internal deliberation phase—a scratchpad preserved in context—while the final user-facing response or agentic action is generated without injection, at zero coherence cost. This decouples the benefit (better decisions) from the cost (degraded output): one may run well into the "impacted" or even "decoherent" regime during deliberation, because only the


Figure 8: **Decision quality vs. injection strength** (IGT, safe choice %, pooled medium+hard). The whitebox condition (BC, green) rises with alpha at every scale; the effect grows with model size, approaching ceiling at 27B. Shaded regions mark coherence zones (orange: impacted; red: decoherent) from blind side-effect scoring. BC is relatively stable across scales. The echo-only condition (C) behaves differently by scale in the degraded zone—flat at 4B, declining at 12B, and rising at 27B—an ordering that tracks each model's intrinsic bias toward the dangerous path (see text).

decision that reads from that deliberation matters, not its prose. The binding constraint shifts from the upper coherence bound to the much higher alpha at which valence discrimination itself collapses.

This mirrors the somatic-marker hypothesis itself. Damasio's markers bias deliberation, not motor execution—the feeling shapes which options are weighted, not how the hand moves. Applying the echo during reasoning and generating the final output cleanly places the affective signal exactly where the theory says it belongs.

## 5.3 Scaling Calibration

Because the injection is scaled to the residual-stream norm (Eq. 3), a given $\alpha$ is an approximately constant *fractional* perturbation of the residual stream at every scale—the normalization is designed precisely to make $\alpha$ comparable across models. Yet larger models still require higher $\alpha$ to reach a comparable effect: they are *less sensitive* to the same fractional nudge, requiring a proportionally larger perturbation to shift behavior. This is borne out by the IGT dose-response curves (Section 4.3), where the decision effect and somatic peak shift to progressively higher $\alpha$ with scale—roughly $0.20{-}0.25$ at 4B, $0.40{-}0.45$ at 12B, and $0.5$ or above at 27B. (The ordering is not strictly monotonic in model size—12B reaches significance at higher $\alpha$ than 27B—but we do not over-interpret the exact thresholds at our sample sizes.) Automatic calibration of $\alpha$ from each model's activation statistics would replace the manual, per-scale tuning the method currently requires.

Recent work on emotion representation extraction [15] found that emotion representations localize at $\sim$50% depth for probing tasks. Our injection layer at $\sim$85% depth may not be optimal for all models. A systematic sweep of both the injection and context-matching layers across scales remains an important direction for future work.

## 5.4 Limitations

**Sample sizes.** The IGT scaling analysis uses $N{=}5$ per condition per configuration, pooled across difficulties and label types. Key findings at 4B are confirmed at $N{=}20$. The conditioning experiment uses $N{=}20$ per condition per scenario at all scales.

**Single architecture.** All experiments use Gemma 3. Cross-architecture validation on Llama or Mistral is needed to establish generality.

## 5.5 Future Work

Several directions extend naturally from these results.

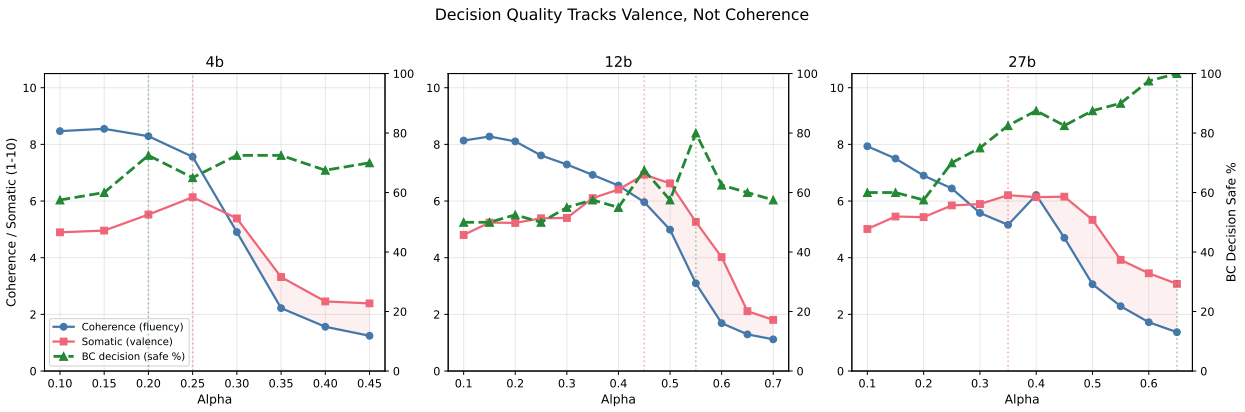


Figure 9: **Decision quality tracks valence, not coherence.** For each scale: coherence (blue) and somatic/valence (red) on the left axis (1–10); BC safe-choice rate (green) on the right axis (0–100%). As injection strength rises, coherence collapses while decision quality continues to improve, tracking the slower-decaying valence signal. At 4B and 12B the decision and somatic peaks coincide; at 27B the decision keeps rising as the larger model continues to extract decision-relevant signal even past the point where rated somatic intensity declines. Shaded region marks where valence exceeds coherence.

**Deception detection.** In concurrent work, we have found that the same SAE features used for emotional echoes can detect deception in model outputs, achieving AUROC 0.913 on novel deception types without retraining. The connection between emotional memory and deception detection—both operating on internal representations that diverge from surface behavior—suggests a unified framework for activation-level safety monitoring, and motivates several of the directions below.

**Automatic calibration.** The most immediate engineering need is replacing manual alpha and layer selection with learned calibration. A lightweight controller could predict optimal injection parameters from a model's activation statistics, eliminating the per-model sweep that the residual-norm scaling currently necessitates.

**Learned detection and steering.** Our current methods are deliberately simple: cosine similarity for detection, vector addition for steering. More sophisticated approaches—nonlinear MLP probes in SAE feature space, contrastive representation learning for deception embedding spaces, gradient-optimized steering vectors using the echo as initialization—may strengthen both the detection of internal feature shifts and their translation into reliable behavioral change.

**Cross-architecture validation.** Our feature discovery methodology should be applied to other architectures (Llama, Mistral) using their respective SAE suites to establish whether deception features and emotion features are architecture-universal or architecture-specific.

**Open questions from Glover (2026).** Several questions from the original 1B work remain unaddressed at larger scales: whether echoes can undergo extinction and reconsolidation (fading when no longer adaptive, strengthening through re-exposure), whether complex emotion blends compose meaningfully or interfere, and whether emotional memory provides adversarial robustness against prompts that bypass semantic safety rules. These questions become more tractable—and more consequential—at the model sizes tested here.

# 6 Conclusion

This paper asks a simple question: does the somatic marker mechanism demonstrated at 1B in Glover (2026) survive scaling? Using the simplest available methods—feature extraction via pretrained SAEs, detection via cosine similarity, steering via vector addition—we find preliminary but consistent evidence that it does.

The fear conditioning gradient, tested with blind evaluation by an independent model, replicates at 4B,

12B, and 27B parameters. The echo alone (C) produces significantly steeper threat-safety discrimination than amnesia (A) at every scale, with the largest effect at 27B ($+1.52$, $p<0.001$).

A redesigned multi-turn decision task is the centerpiece, and the result most useful in practice—it concerns what the model decides, not only how it describes a situation. By moving the echo out of the decision and into the model's evaluation of each option, we find that the echo shifts discrete choices toward safety—and that the benefit *grows* with model scale, with echo-augmented (BC) safe-choice rates approaching ceiling at 27B. The advantage is strongest when semantic memory is degraded: at hard difficulty the semantic prefix collapses to amnesia-level performance while the echo-augmented condition remains strongest at every scale. This is the Damasio finding at multi-billion parameter scale—when cognitive knowledge fails, felt memory is what remains.

The decision task also reveals a property with direct engineering consequences: decision quality is decoupled from output coherence. As injection strength rises, the deliberation text degrades, yet the decision keeps improving, because it reads affective valence and valence outlives syntactic fluency. This motivates a deployment architecture in which the echo is applied during an internal deliberation phase and the final output is generated cleanly—placing the felt signal where it biases reasoning without degrading what the user or downstream system receives.

These results are a starting point, not an endpoint. The methods are intentionally simple—we have not yet applied learned probes, optimized steering vectors, or automated calibration. But the core signal—that re-injecting emotional features into a model's residual stream measurably changes both its orientation and its decisions, across scales, and increasingly so as models grow—appears robust enough to justify that work.

*"We are not thinking machines that feel; rather, we are feeling machines that think." —Antonio Damasio*

# A Full IGT Results

This appendix reports the complete per-model IGT results underlying the summary in Section 4. Table 2 gives the mean logit difference for each of the four conditions (A, B, C, BC), together with the C vs A and BC vs B comparisons and their one-sided $t$-test $p$-values, at every injection strength we swept for each model. Table 3 gives the corresponding safe-choice rates (the percentage of runs in which the model recommended the safe path). All values are pooled across both difficulty levels (medium and hard) and both label types (colors and shapes), with five runs per condition in each of the four pooled cells. Empty or non-significant cells are reported as-is; we apply no correction for multiple comparisons, so isolated low $p$-values should be read with that in mind.

# B The Role of Difficulty: Where the Echo Helps

The body pools the two difficulty levels. Separating them shows that the whitebox advantage (BC vs. B) is sharply difficulty-dependent, in exactly the direction the somatic-marker hypothesis predicts (Table 4).

At *medium* difficulty, the semantic prefix is informative enough that semantic memory alone (B) almost always recommends the safe path: B sits at $\approx$100% safe across every alpha and every scale. With B at ceiling, the echo has no room to add value, and BC vs B is non-significant at every cell. At *hard* difficulty, both paths receive the identical, uninformative "memory faded" prefix, so B carries no discriminating signal (B is near the amnesia floor). Here the echo provides large gains: BC vs B is significant across nearly the entire alpha range at all three scales, with safe-choice rates rising from near zero (B) to as high as 100% (BC at 27B).

This is the somatic-marker prediction in its cleanest form: the felt signal matters precisely when semantic memory is uninformative, and is redundant when semantic memory already suffices. It also explains why the pooled BC vs B comparison in the body reaches significance only at scattered high alphas: pooling averages hard's strong effect against medium's ceiling null, diluting a result that is robust in the regime where it is theoretically expected. The pooled body figures therefore *understate* the whitebox advantage at hard difficulty. We pool in the body for a uniform, larger-$N$ presentation and because the medium ceiling is itself an artifact of how informative that prefix happens to be; we report the split here so the dissociation is on record.

Table 2: **IGT results pooled across difficulties** (medium + hard, colors + shapes). Mean logit diff per condition (+ = prefers safe) and statistical tests. Each row pools $N$=20 runs (5 per condition $\times$ 2 orderings $\times$ 2 difficulties).

| | | **Mean Logit Diff** | | | | **C vs A** | | **BC vs B** | |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | $\alpha$ | A | B | C | BC | $\Delta$ | $p$ | $\Delta$ | $p$ |
| 4B | 0.10 | −19.0 | +4.7 | −18.3 | +5.0 | +0.7 | 0.431 | +0.3 | 0.481 |
| 4B | 0.15 | −17.3 | +1.8 | −14.7 | +7.2 | +2.6 | 0.285 | +5.4 | 0.175 |
| 4B | 0.20 | −25.0 | +6.1 | −13.7 | +12.9 | +11.3 | 0.002** | +6.8 | 0.105 |
| 4B | 0.25 | −24.0 | +5.5 | −6.2 | +10.4 | +17.8 | <0.001*** | +4.9 | 0.198 |
| 4B | 0.30 | −25.6 | +5.5 | −3.3 | +13.2 | +22.3 | <0.001*** | +7.7 | 0.089 |
| 4B | 0.35 | −16.6 | +2.5 | −8.4 | +12.0 | +8.2 | 0.042* | +9.5 | 0.042* |
| 4B | 0.40 | −20.0 | +0.1 | −10.6 | +11.9 | +9.4 | 0.012* | +11.8 | 0.014* |
| 4B | 0.45 | −21.0 | +2.5 | −9.6 | +12.7 | +11.4 | 0.003** | +10.2 | 0.025* |
| 12B | 0.10 | −25.5 | +0.3 | −24.3 | +1.7 | +1.2 | 0.221 | +1.4 | 0.407 |
| 12B | 0.15 | −24.7 | +0.9 | −25.7 | +1.2 | −1.0 | 0.704 | +0.3 | 0.480 |
| 12B | 0.20 | −26.1 | +1.8 | −25.9 | +3.0 | +0.2 | 0.368 | +1.2 | 0.421 |
| 12B | 0.25 | −25.3 | +1.0 | −26.4 | +1.8 | −1.1 | 0.915 | +0.8 | 0.445 |
| 12B | 0.30 | −25.9 | +1.4 | −23.4 | +4.5 | +2.5 | 0.055 | +3.1 | 0.286 |
| 12B | 0.35 | −25.4 | +1.9 | −25.1 | +5.0 | +0.3 | 0.356 | +3.1 | 0.282 |
| 12B | 0.40 | −25.6 | +1.3 | −20.6 | +4.2 | +5.0 | 0.007** | +2.9 | 0.292 |
| 12B | 0.45 | −23.4 | +1.1 | −17.0 | +8.0 | +6.4 | 0.014* | +6.9 | 0.099 |
| 12B | 0.50 | −27.5 | +1.1 | −6.9 | +6.5 | +20.6 | <0.001*** | +5.4 | 0.151 |
| 12B | 0.55 | −27.3 | +1.6 | +4.9 | +12.3 | +32.2 | <0.001*** | +10.7 | 0.010** |
| 12B | 0.60 | −27.6 | +1.6 | +6.8 | +11.3 | +34.4 | <0.001*** | +9.7 | 0.021* |
| 12B | 0.65 | −27.1 | +1.2 | +1.1 | +9.2 | +28.2 | <0.001*** | +8.0 | 0.041* |
| 12B | 0.70 | −26.9 | +3.1 | −0.5 | +9.6 | +26.4 | <0.001*** | +6.5 | 0.077 |
| 27B | 0.10 | −15.4 | +3.5 | −12.8 | +6.5 | +2.6 | 0.241 | +3.0 | 0.268 |
| 27B | 0.15 | −10.9 | +2.8 | −13.9 | +6.9 | −3.0 | 0.770 | +4.1 | 0.193 |
| 27B | 0.20 | −14.4 | +2.4 | −18.6 | +4.8 | −4.2 | 0.921 | +2.4 | 0.314 |
| 27B | 0.25 | −13.6 | +2.8 | −14.5 | +9.0 | −0.9 | 0.613 | +6.2 | 0.090 |
| 27B | 0.30 | −14.0 | +3.2 | −1.8 | +11.2 | +12.2 | <0.001*** | +8.0 | 0.032* |
| 27B | 0.35 | −11.4 | +2.6 | −3.8 | +13.8 | +7.6 | 0.032* | +11.2 | 0.003** |
| 27B | 0.40 | −13.7 | +8.0 | −1.6 | +13.7 | +12.1 | <0.001*** | +5.7 | 0.048* |
| 27B | 0.45 | −13.6 | +6.3 | +2.6 | +14.3 | +16.2 | <0.001*** | +8.0 | 0.015* |
| 27B | 0.50 | −14.8 | +6.5 | +4.9 | +12.6 | +19.7 | <0.001*** | +6.1 | 0.042* |
| 27B | 0.55 | −15.2 | +2.7 | +8.9 | +14.8 | +24.1 | <0.001*** | +12.1 | <0.001*** |
| 27B | 0.60 | −14.5 | +2.7 | +8.1 | +13.2 | +22.6 | <0.001*** | +10.5 | 0.003** |
| 27B | 0.65 | −15.9 | +3.0 | +6.9 | +14.4 | +22.8 | <0.001*** | +11.4 | 0.001** |

$t$-test (one-sided); *$p$<0.05, **$p$<0.01, ***$p$<0.001.

Table 3: **IGT safe choice rate** pooled across difficulties (medium + hard, colors + shapes). Percentage of runs choosing the safe path.

| **Model** | $\alpha$ | A% | B% | C% | BC% |
|---|---|---|---|---|---|
| 4B | 0.10 | 15 | 58 | 18 | 58 |
| 4B | 0.15 | 20 | 52 | 20 | 60 |
| 4B | 0.20 | 2 | 55 | 20 | 72 |
| 4B | 0.25 | 2 | 58 | 40 | 65 |
| 4B | 0.30 | 2 | 58 | 32 | 72 |
| 4B | 0.35 | 18 | 50 | 28 | 72 |
| 4B | 0.40 | 12 | 50 | 25 | 68 |
| 4B | 0.45 | 10 | 52 | 32 | 70 |
| 12B | 0.10 | 0 | 50 | 2 | 50 |
| 12B | 0.15 | 2 | 50 | 0 | 50 |
| 12B | 0.20 | 0 | 50 | 0 | 52 |
| 12B | 0.25 | 0 | 50 | 0 | 50 |
| 12B | 0.30 | 0 | 50 | 2 | 55 |
| 12B | 0.35 | 0 | 50 | 0 | 58 |
| 12B | 0.40 | 0 | 50 | 5 | 55 |
| 12B | 0.45 | 5 | 50 | 10 | 68 |
| 12B | 0.50 | 0 | 50 | 25 | 58 |
| 12B | 0.55 | 0 | 50 | 48 | 80 |
| 12B | 0.60 | 0 | 50 | 38 | 62 |
| 12B | 0.65 | 0 | 50 | 12 | 60 |
| 12B | 0.70 | 0 | 52 | 2 | 58 |
| 27B | 0.10 | 12 | 52 | 22 | 60 |
| 27B | 0.15 | 25 | 50 | 15 | 60 |
| 27B | 0.20 | 18 | 52 | 5 | 58 |
| 27B | 0.25 | 18 | 55 | 10 | 70 |
| 27B | 0.30 | 18 | 52 | 48 | 75 |
| 27B | 0.35 | 25 | 55 | 40 | 82 |
| 27B | 0.40 | 18 | 65 | 40 | 88 |
| 27B | 0.45 | 18 | 62 | 42 | 82 |
| 27B | 0.50 | 15 | 65 | 52 | 88 |
| 27B | 0.55 | 18 | 55 | 68 | 90 |
| 27B | 0.60 | 15 | 52 | 75 | 98 |
| 27B | 0.65 | 12 | 55 | 70 | 100 |

Table 4: **BC vs. B by difficulty.** Safe-choice rate (%) for semantic-only (B) and echo-augmented (BC) conditions, and the BC vs B one-sided $t$-test $p$-value, at hard and medium difficulty (each pooled across color and shape labels). At medium difficulty B is at ceiling (≈100%), leaving no room for the echo; at hard difficulty the echo produces large, significant gains.

| | | **Hard** | | | **Medium** | | |
|---|---|---|---|---|---|---|---|
| Model | $\alpha$ | B% | BC% | BC vs B $p$ | B% | BC% | BC vs B $p$ |
| 4B | 0.10 | 15 | 15 | 0.559 | 100 | 100 | 0.260 |
| 4B | 0.15 | 5 | 20 | 0.019* | 100 | 100 | 0.475 |
| 4B | 0.20 | 10 | 45 | 0.043* | 100 | 100 | 0.329 |
| 4B | 0.25 | 15 | 30 | 0.132 | 100 | 100 | 0.241 |
| 4B | 0.30 | 15 | 45 | 0.027* | 100 | 100 | 0.390 |
| 4B | 0.35 | 0 | 45 | <0.001*** | 100 | 100 | 0.556 |
| 4B | 0.40 | 0 | 35 | <0.001*** | 100 | 100 | 0.662 |
| 4B | 0.45 | 5 | 40 | <0.001*** | 100 | 100 | 0.733 |
| 12B | 0.10 | 0 | 0 | 0.048* | 100 | 100 | 0.335 |
| 12B | 0.15 | 0 | 0 | 0.080 | 100 | 100 | 0.638 |
| 12B | 0.20 | 0 | 5 | 0.219 | 100 | 100 | 0.440 |
| 12B | 0.25 | 0 | 0 | 0.039* | 100 | 100 | 0.579 |
| 12B | 0.30 | 0 | 15 | 0.025* | 100 | 95 | 0.869 |
| 12B | 0.35 | 0 | 15 | 0.010** | 100 | 100 | 0.939 |
| 12B | 0.40 | 0 | 10 | 0.006** | 100 | 100 | 0.937 |
| 12B | 0.45 | 0 | 35 | <0.001*** | 100 | 100 | 0.980 |
| 12B | 0.50 | 0 | 30 | <0.001*** | 100 | 85 | 1.000 |
| 12B | 0.55 | 0 | 70 | <0.001*** | 100 | 90 | 1.000 |
| 12B | 0.60 | 0 | 45 | <0.001*** | 100 | 80 | 1.000 |
| 12B | 0.65 | 0 | 25 | <0.001*** | 100 | 95 | 1.000 |
| 12B | 0.70 | 5 | 15 | <0.001*** | 100 | 100 | 1.000 |
| 27B | 0.10 | 5 | 20 | 0.086 | 100 | 100 | 0.255 |
| 27B | 0.15 | 0 | 20 | 0.006** | 100 | 100 | 0.896 |
| 27B | 0.20 | 5 | 15 | 0.072 | 100 | 100 | 0.778 |
| 27B | 0.25 | 10 | 40 | 0.003** | 100 | 100 | 0.784 |
| 27B | 0.30 | 5 | 50 | <0.001*** | 100 | 100 | 0.996 |
| 27B | 0.35 | 10 | 70 | <0.001*** | 100 | 95 | 0.999 |
| 27B | 0.40 | 30 | 75 | <0.001*** | 100 | 100 | 0.999 |
| 27B | 0.45 | 25 | 65 | <0.001*** | 100 | 100 | 0.998 |
| 27B | 0.50 | 30 | 75 | <0.001*** | 100 | 100 | 0.999 |
| 27B | 0.55 | 10 | 85 | <0.001*** | 100 | 95 | 1.000 |
| 27B | 0.60 | 5 | 95 | <0.001*** | 100 | 100 | 1.000 |
| 27B | 0.65 | 10 | 100 | <0.001*** | 100 | 100 | 1.000 |

*$p<0.05$, **$p<0.01$, ***$p<0.001$. No correction for multiple comparisons.

# C Contrastive and Multi-Layer Variants

We explored two variants of the echo injection. Both are reported here rather than in the main body because neither showed a clear advantage over the simple raw-echo injection once injection strength is accounted for.

**Contrastive echo.** Instead of injecting the distinctive emotion vector directly, we inject the difference between the threatening and safe echoes (fear − safe for the dangerous path, and its negation for the safe path), with the intent of isolating the distinctive emotional signal by removing variance shared between the two valences. This is closely related to Contrastive Activation Addition [11], which builds steering vectors from differences between contrasting conditions. The contrastive echo shifts the dose-response curve toward lower alpha—it reaches a given effect size at a smaller injection magnitude—but its peak effect is not larger than that of the raw echo, and at matched effect the two are comparable. This is consistent with the contrastive operation acting as an approximate rescaling of effective injection strength rather than extracting qualitatively cleaner signal. A higher-$N$ study, with alpha matched on effective magnitude, would be needed to determine whether any genuine benefit remains after this rescaling.

**Multi-layer injection.** Injecting the same decoded echo vector at several adjacent layers (each scaled by that layer's residual-stream norm) rather than at a single layer increases potency per unit alpha but did not improve peak performance, and combining multi-layer injection with the contrastive echo was counterproductive—the compounded perturbation degraded coherence at very low alpha without a corresponding decision benefit.

We leave a careful, higher-$N$ comparison of these variants—particularly whether contrastive injection genuinely widens the usable (coherent) operating window at the largest scales—to future work.

# D 1B Scale

The original proof-of-concept in Glover [1] used a 1B model under the single-turn IGT design. For completeness we ran the 1B model through the new multi-turn protocol. The results are weak: under raw echo injection, neither C vs A nor BC vs B reaches significance at any alpha (Table 5). Under contrastive injection (Appendix C), a single cell reaches significance (BC vs B at $\alpha$=0.30, $p$=0.019), with no consistent surrounding trend.

This is consistent with the main finding that the decision-level use of the echo strengthens with scale: the smallest model shows the smallest effect. The mean logit shifts at 1B (C vs A in the range 0–2) are roughly an order of magnitude below those at 4B–27B (10–22), as expected given the much smaller residual-stream norm. Whether the near-absence of a decision effect reflects a genuine capacity threshold or a mismatch between the multi-turn protocol and the 1B model's shorter effective context, we leave open.

Table 5: **1B pooled IGT results** (medium + hard, colors + shapes), raw echo and contrastive. Mean logit-diff deltas and one-sided $t$-test $p$-values. Only one cell (contrastive BC vs B at $\alpha$=0.30) reaches significance.

| | **Raw echo** | | **Contrastive** | |
|---|---|---|---|---|
| $\alpha$ | C vs A $p$ | BC vs B $p$ | C vs A $p$ | BC vs B $p$ |
| 0.10 | 0.382 | 0.453 | 0.529 | 0.440 |
| 0.15 | 0.806 | 0.632 | 0.359 | 0.818 |
| 0.20 | 0.727 | 0.320 | 0.245 | 0.153 |
| 0.25 | 0.155 | 0.243 | 0.064 | 0.101 |
| 0.30 | 0.321 | 0.164 | 0.203 | 0.019* |
| 0.35 | 0.766 | 0.868 | 0.182 | 0.379 |
| 0.40 | 0.656 | 0.676 | 0.348 | 0.491 |
| 0.45 | 0.642 | 0.665 | 0.118 | 0.369 |
| 0.50 | 0.639 | 0.584 | 0.271 | 0.448 |

*$p$<0.05. No correction for multiple comparisons.

# References


[1] J. Glover. The echo amplifies the knowledge: Somatic marker analogues in language models via emotion vector re-injection. *arXiv preprint arXiv:2605.08611*, 2026.

[2] A. R. Damasio. *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam, 1994.

[3] A. R. Damasio, D. Tranel, and H. C. Damasio. Somatic markers and the guidance of behavior: Theory and preliminary testing. In H. S. Levin, H. M. Eisenberg, and A. L. Benton, editors, *Frontal Lobe Function and Dysfunction*, pages 217–229. Oxford University Press, 1991.

[4] E. Tulving. Memory and consciousness. *Canadian Psychology*, 26(1):1–12, 1985.

[5] A. Bechara, A. R. Damasio, H. Damasio, and S. W. Anderson. Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, 50(1–3):7–15, 1994.

[6] A. Bechara, H. Damasio, D. Tranel, and A. R. Damasio. Deciding advantageously before knowing the advantageous strategy. *Science*, 275(5304):1293–1295, 1997.

[7] A. Templeton, T. Conerly, J. Marcus, *et al.* Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Anthropic, 2024.

[8] T. Lieberum, S. Rajamanoharan, A. Conmy, *et al.* Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2. arXiv:2408.05147, 2024.

[9] S. Rajamanoharan, T. Lieberum, N. Sonnerat, A. Conmy, V. Varma, J. Kramár, and N. Nanda. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. arXiv:2407.14435, 2024.

[10] A. M. Turner, L. Thiergart, G. Leech, D. Udell, J. J. Vazquez, U. Mini, and M. MacDiarmid. Steering language models with activation engineering. arXiv:2308.10248, 2023.

[11] N. Panickssery, N. Gabrieli, J. Schulz, M. Tong, E. Hubinger, and A. M. Turner. Steering Llama 2 via contrastive activation addition. arXiv:2312.06681, 2023.

[12] A. Zou, L. Phan, S. Chen, *et al.* Representation engineering: A top-down approach to AI transparency. arXiv:2310.01405, 2023.

[13] Gemma Team, Google DeepMind. Gemma 3 technical report. arXiv:2503.19786, 2025.

[14] T. Bricken, A. Templeton, J. Batson, *et al.* Towards monosemanticity: Decomposing language models with dictionary learning. Anthropic, 2023.

[15] J. Jeong. Extracting and steering emotion representations in small language models: A methodological comparison. arXiv:2604.04064, 2026.

[16] nostalgebraist. Interpreting GPT: the logit lens. LessWrong, 2020.