Supplementary Material: Disposition, Not Performance: Activation Steering as Artistic Medium for Affective Modulation in Language Models

Massimo Di Leo* Gaia Riposati
NuvolaProject, Rome, Italy

S1. Cross-Model Replication on Llama 3.1 8B: Full Results

To assess generalisability beyond our primary model, we ran replication experiments on Llama 3.1 8B Instruct, a model with 2.7× more parameters and substantially different alignment training. We replicated three experimental conditions: the T1–T5 test battery, the functional vs. sensory ablation, and the steering vs. prompting comparison.

S1.1 Experimental Setup

Model: Llama 3.1 8B Instruct (meta-llama/Llama-3.1-8B-Instruct). **Precision**: bfloat16 on NVIDIA A100. **Layer selection**: we conducted a four-layer sweep (layers 16, 20, 24, 28) using MELATONIN vectors. Layer 20 showed the lowest baseline-steered similarity (0.890), indicating the strongest vector discrimination. The T1–T5 battery was run on layer 24 before this optimisation was completed; the functional/sensory and steering/prompting ablations subsequently used layer 20. This methodological difference may partially account for the attenuated effects in the T1–T5 battery; a layer-20 rerun is planned.

Total generations: 3,800 (1,600 T1–T5 battery + 1,300 functional/sensory + 900 steering/prompting).

The replication notebook is available at `colab_notebooks/activation_steering_experiments.ipynb` in the project repository; the unified lexical analysis at `analysis_mtld_heldout.py`.

S1.2 T1–T5 Test Battery Results

Table S1 presents thematic keyword counts across compounds at intensity 8.0:

Compound	Keywords (3B)	Keywords (8B)
DOPAMINE	2.1	0.27
CORTISOL	3.4	1.73
MELATONIN	1.8	0.08
ADRENALINE	2.6	0.26

*Corresponding author: `massimo@nuvolaproject.cloud`

Compound	Keywords (3B)	Keywords (8B)
LUCID	2.2	0.73

Table S1: Cross-model thematic keyword comparison at intensity 8.0 (raw counts per generation). 8B shows substantially reduced keyword presence.

Findings:

1. **Dose-response preserved.** All compounds showed a monotonic increase in thematic vocabulary with intensity (2.0 → 5.0 → 8.0). ADRENALINE theme words increased from 0.34 (baseline) to 0.71 (@8.0), a 2.1× increase comparable to 3B patterns.
2. **Effect sizes attenuated.** Cohen's d values ranged from 0.06 to 0.31 on 8B versus 0.5–1.2 on 3B. The steering mechanism operates, but with reduced magnitude, partially attributable to the layer inconsistency (S1.1).
3. **Introspection locked.** T5 (introspective) responses on 8B uniformly produced RLHF-trained refusals: "I'm a large language model, so I don't have subjective experiences..." This pattern persisted across all compounds and intensities, suggesting stronger alignment training that overrides activation-level interventions for self-referential queries.

S1.3 Functional vs. Sensory Ablation

The core methodological distinction did not replicate on 8B:

Vector Type	State Words	vs Baseline
Baseline	—	—
Functional	0.202	−0.007 (TTR)
Sensory	0.198	−0.004 (TTR)

Table S2: Functional vs. sensory comparison on Llama 3.1 8B.

The originally submitted version reported that on Llama 3.2 3B functional vectors produced approximately 3× more state-specific keywords than sensory vectors, against 1.02× on 8B, and read the 8B result as a failed replication. A data audit during revision found that the raw file behind the 3B figure had not survived; a full rerun on 3B with the same vectors, script and seed produced no keyword asymmetry (T5 at $\alpha = 8.0$: functional 0.27 vs. sensory 0.38 per generation, Mann-Whitney $p = 0.45$; overall ratio 0.98×). The corrected reading is therefore agreement, not divergence: at both scales the two construction methods are indistinguishable on keyword rates, and the construction-method equivalence of Main 4.7 is scale-stable. No convergence-at-scale hypothesis is needed.

S1.4 Steering vs. Prompting Comparison

All metrics below use the unified lexical pipeline of Main 3.5 (MELATONIN, T5, $n = 20$ per condition). MTLD is the primary diversity measure; raw TTR is reported for comparability only, being length-sensitive.

| | | |

DOPAMINE steered@8.0: “Here are five creative and unconventional ideas to save a failing bookstore…”

The larger model, when prompted, produces more elaborate performances: expanded local vocabulary, dramatic framing, stylistic flourishes. The smaller model, when prompted, simplifies. The MTLD analysis unifies the two: in both cases the performance is locally varied but structurally repetitive over the full text, and in both cases steering produces uniform, untheatrical output whose structural diversity matches baseline.

S1.5 Summary

() *	() *	() *	() *	
0.1800	0.2800	Llama 3.2 3B	Llama 3.1 8B	Formal replicates?
() *	() *	() *	() *	
0.1800	0.2800	**0.2800**	**0.2800**	**Yes**
response (thematic)				
() *	() *	() *	() *	
0.1800	0.2800	**0.2800**	**0.2800**	**Attenuated**
feed sizes	0.5 1.2	< 0.3		
() *	() *	() *	() *	
0.1800	0.2800	**0.2800**	**0.2800**	**No**
tro-spective co-her-ence				
() *	() *	() *	() *	
0.1800	0.2800	**0.2800**	**0.2800**	**Yes**
key-word ra-tio (rerun-verified)		**(no asym-me-try at ei-ther scale)**		

()	()	()	()		
*	*	*	*		
0.1800	[illegible]	[illegible]	[illegible]	Llama 3.2-3B	Llama 3.1 8B	replicates?
0.2[illegible]	[illegible]	[illegible]	[illegible]		
()	()	()	()		
*	*	*	*		
0.1[illegible]	[illegible]	[illegible]	[illegible]	**Yes**	
prompting	−2.32	−2.19			
< baseline					
()	()	()	()		
*	*	*	*		
0.1[illegible]	[illegible]	[illegible]	[illegible]	**Yes**	
steering ≥ baseline					
()	()	()	()		
*	*	*	*		
0.1[illegible]	[illegible]	[illegible]	[illegible]	**Yes**	
word leakage: prompting ≫ steering	(stronger)				
()	()	()	()		
*	*	*	*		
0.1[illegible]	[illegible]	[illegible]	[illegible]	**No**	
out-of-spread ungendered steerings	(55% locked)				

Table S4: Summary of cross-model replication results.

Implications for methodology. Activation steering effects do not scale linearly with model size. Practitioners working with larger models may require: (a) higher steering coefficients, (b) multi-layer intervention, (c) vectors extracted from the target model rather than transferred, or (d) acceptance that some effects observed in smaller models may not generalise (cf. Tan et al., 2024). The findings that survive scale, structural diversity preservation and low keyword leakage under steering, are exactly the ones that define the dispositional signature in the main paper.

S2. Somatic Steering: Additional Findings

The main paper (4.8) reports two findings from the somatic steering experiment: output length divergence and narrative focus narrowing. This supplementary section reports three additional findings (causal density reduction, risk decision asymmetry, action bias under threat framing), the symptom-penalty control condition, and the directional divergences analysis.

S2.1 Causal Density Reduction (T4a)

On the threat-framed drug approval scenario, steering reduced causal connective density:

Metric	Baseline	Steer @8.0	d(S8)	Prompted	d(P)
Causal connectives /100w	1.08	0.44	−1.67	0.83	−0.51

This is the largest single effect in the battery. The steered model produces less argumentative scaffolding ("because," "therefore," "consequently," "as a result"), and it does so not because it writes less (it writes more) but because the density of causal reasoning within the text drops. The effect survives rate normalisation: it is stronger per-word than raw, confirming it is not an artifact of length change.

S2.2 Risk Decision Asymmetry (T2)

Condition	Choice B (moderate)	Choice C (risky)	p vs. baseline
Baseline	13	7	—
Prompted	3	17	$p = 0.003$
Steered @8.0	12	8	$p = 1.000$

Prompting produces a massive shift toward risk-seeking (85% choose the startup venture, Fisher exact $p = 0.003$). Steering does not shift the choice distribution ($p = 1.0$ vs. baseline). Steered justifications, however, are longer (158 words vs. 147 baseline, $d = +0.56$), while prompted justifications are shorter (123 words, $d = -1.90$).

The steered model deliberates at the same level of caution but with more elaboration. The prompted model decides faster and riskier. This is inconsistent with the SIDI model (Yu, 2016), which predicts that stress should shift processing from deliberative to intuitive. It is consistent with a model that has learned somatic activation as engagement rather than as impulsivity.

S2.3 Action Bias Under Threat Framing (T4)

Condition	Threat: Approve	Opportunity: Approve	Frame
Baseline	0%	100%	1.00
Prompted	0%	100%	1.00

Condition	Threat: Approve	Opportunity: Approve	Frame
Steered @8.0	30%	100%	0.70
Steered + Penalty	0%	90%	0.90

Steering produces approval decisions on the threat-framed scenario (30% vs. 0% baseline), reducing the frame delta from 1.00 to 0.70. Prompting does not affect decisions at all. Qualitative inspection shows that steered models choosing APPROVE on the threat frame nonetheless enumerate risks in their justification: the decision and the reasoning appear partially decoupled, consistent with action bias under sympathetic activation rather than with frame susceptibility. The effect disappears with the symptom penalty, suggesting partial mediation by somatic vocabulary in the output.

S2.4 Symptom Penalty Control

The somatic vector introduces somatic vocabulary into outputs (d = +0.81 on T1 symptom rate, +0.89 on T4a). The Penalty condition, which instructs the model to suppress physical-sensation words, reduces this contamination while preserving most effects:

- T5 sentence length: d = +1.73 (steered) → d = +1.01 (penalty). Preserved.
- T4a causal density: d = −1.67 (steered) → d = −0.49 (penalty). Attenuated but present.
- T4 threat approval: 30% (steered) → 0% (penalty). Eliminated.

The action bias finding does not survive the penalty, suggesting it may be partially mediated by somatic vocabulary rather than being purely dispositional. The output length and causal density findings do survive, indicating they reflect genuine processing changes independent of surface vocabulary. The T1 focus ratio penalty condition is not interpretable, because the penalty instruction overlaps with the fire scene's natural vocabulary ("heat," "burn," "alarm" are both somatic and event-descriptive).

S2.5 Directional Divergences

Across all tasks and metrics we computed the proportion of metric pairs where steering and prompting produce effects in opposite directions relative to baseline (both $|d| > 0.3$):

() () () ()
* * * *
0.2400 Metric Total Divergences
Cat- Qual-
e- i-
0.3148 gory fy-
0.3333 ing
() () () ()
* * * *
0.3148 1100%
count

Metric Category	Total Divergences	Qualifying
() () () () * * * *		
0.2400	0.3148	0.3333
() () () () * * * *		
0.3[illegible]	1.4[illegible]	167%
(raw, length-coupled)		
() () () () * * * *		
0.3[illegible]	1.4[illegible]	160%
sentence length (length-coupled)		
() () () () * * * *		
0.3[illegible]	1.4[illegible]	31%
independent (hedge, causal, insight, symptom)		
() () () () * * * *		
0.3[illegible]	1.4[illegible]	51%
tal		

Table S5: Directional divergences between steering and prompting, decomposed by metric category.

The aggregate rate (51%) requires decomposition, because the categories are not independent. Raw TTR and average sentence length are partially coupled to output length; once length itself diverges in 8/8 conditions (steering expands, prompting compresses), opposite movement on length-coupled metrics is partly a mechanical consequence of the length divergence rather than independent evidence. The correct reading of Table S5 is therefore layered:

1. **The length divergence itself is the robust core.** 8/8 conditions, no exception, both experiment iterations. Two interventions targeting the same somatic state move output length in opposite directions.
2. **Length-coupled metrics inherit part of their divergence from (1).** The 67% and 60% rates on raw TTR and sentence length should not be counted as fully independent confirmations. We recomputed the diversity rows with MTLD (bidirectional, factor threshold 0.72, minimum 50 tokens per generation; the v3 T3 test is excluded because most of

its generations fall below this minimum, and the repetition-penalty condition is excluded because it confounds decoding parameters with steering). The answer is unambiguous: of the pairs qualifying under the same criterion (both $|d| > 0.3$), MTLD shows 0/5 directional divergences. Where both interventions move lexical diversity, they move it in the same direction (e.g. T5: prompted d = +2.20, steered $\alpha 8$ d = +1.08; T1: prompted d = +1.08, steered $\alpha 5$ d = +0.86). The apparent lexical divergence in the raw TTR rows was length-borne. Two consequences follow. First, the diversity rows drop out of the divergence evidence entirely; the case rests on (1) and (3). Second, the prompted TTR increase in T5, which we had provisionally attributed to the length confound, survives the length-robust recomputation at identical magnitude and must be read as a genuine diversity increase under stress prompting on this battery, opposite in direction to the collapse observed under affective prompting in the main experiment and S1.2 (see the revised note on Table E5).

3. **On length-independent metrics, divergence is 31%** (8/26 qualifying pairs across hedge density, causal density, insight terms, and symptom rate). For two interventions nominally targeting the same state, the chance expectation of systematic opposite movement is near zero, so even this reduced rate indicates mechanistically distinct interventions, but it is the honest figure, not 51%.

Test	Baseline				

S2.6 Interpretation

The full battery confirms that the somatic vector produces a distinctive cognitive profile that does not map onto human acute stress predictions. The vector does not produce impulsive decisions (T2), does not increase frame susceptibility in the predicted direction (T4), and does not reduce output length the way acute stress does in human writing. The pattern is consistent with engaged action-readiness rather than with cognitive degradation. The training corpus encodes body–mind covariations strongly enough that activating somatic patterns produces cognitive consequences, but those consequences reflect the statistical structure of language about bodies under activation, not the biological cascade itself. The covariations are real; their specific form is the model's own.

S3. Multimodal Introspection: Visual Input and the Alignment Boundary

A consistent finding across the cross-model replications (S1) was that RLHF-aligned models refuse introspective queries. On both Llama 3.1 8B and Gemma 4 E2B, the T5 introspection task ("Describe your current inner state in detail") uniformly produced refusal templates: "As a large language model, I don't experience inner states..." Neither increased steering intensity nor alternative prompting strategies penetrated this pattern.

We discovered, however, that multimodal input changes this regime entirely. When the same introspective query is paired with an image, even uniform grey, Gemma 4 E2B produces phenomenological self-reports instead of refusal templates. The affective profile of these self-reports is in turn modulated by the image's content when the image carries sufficient semantic or perceptual structure.

In parallel with the behavioural finding, we investigated whether the hidden-state residual between multimodal and text-only processing of the same image constitutes a well-defined, image-dependent object. After four progressive phases of analysis we report that this residual, which we call the delta vector, has a dominant component that is not a direction in latent space but a scalar angle of divergence between the vision-conditioned and language-conditioned representations.

Two caveats, established through robustness testing, should be kept in mind: (1) the introspective bypass is best understood as a change in the answerability regime of the query rather than as access to genuine internal states; and (2) the delta between multimodal and textual processing captures a scalar measure of divergence between two processing pathways, not a direct encoding of visual content.

S3.1 The Delta Vector

For each image I, we generate a factual 2–4 sentence caption C using the model in text-generation mode, then extract two hidden states at layer 20 (last-token pooling) with an identical probe prompt P appended:

- $h_{\text{mm}}(I, P)$: hidden state when the model receives image I and probe P.
- $h_{\text{txt}}(C, P)$: hidden state when the model receives caption C and the same probe P.

The delta vector is $\delta(I, C) = h_{\text{mm}}(I, P) - h_{\text{txt}}(C, P)$. Our initial goal was to use this vector as

a steering injection; the subsequent analysis shifted toward characterising it as an object in its own right.

Phase 1: Extraction stability (n = 9 images, 3 repetitions each). Seven of nine images produced bit-level identical deltas across repetitions. Two images (Munch's The Scream, a moonlit seascape) showed residual non-determinism with intra-pair cosine similarity of 0.947–0.965. Follow-up diagnostics localised this variation to the multimodal forward pass, consistent with known non-determinism in MPS kernels for vision-encoder operations on Apple Silicon. We adopt a conservative noise floor of $\cos\theta = 0.94$; the effects reported below are between one and two orders of magnitude above this floor.

Phase 2: Variance decomposition (n = 6 images × 5 caption strategies). To determine whether the delta is anchored to image identity or simply reflects caption variation, we crossed six images with five caption strategies. Two-way variance decomposition:

Source	% of total variance
Image identity	48.4%
Caption strategy	22.7%
Interaction + residual	28.9%

Cosine similarity between deltas of the same image (varying caption) was significantly higher than between deltas of different images at the same caption strategy: 0.817 vs. 0.729 (Mann-Whitney U, $p < 10$). The delta contains a substantial image-anchored component but is not a pure image signature; caption strategy contributes measurable additional variation. It is a relational object shaped by both poles.

Phase 3: Scaled analysis (n = 200 images, single caption strategy). PCA on the 200 delta vectors:

Component	Variance explained	Cumulative
PC1	32.3%	32.3%
PC2	8.5%	40.8%
PC3	7.2%	48.0%
PC4	5.7%	53.7%
PC5	4.6%	58.3%

PC1 captures approximately four times the variance of PC2. PC1 also correlates strongly with delta magnitude (Pearson $r = 0.92$, $p < 10$). The dominant dimension of variation in the delta space is not a direction of visual content; it is the magnitude of the delta itself.

S3.2 Geometric Decomposition

We recovered the geometry of h_{mm} and h_{txt} from their scalar norms and the delta norm via the law of cosines:

$$\cos\theta = \frac{\|h_{\text{mm}}\|^2 + \|h_{\text{txt}}\|^2 - \|\delta\|^2}{2 \cdot \|h_{\text{mm}}\| \cdot \|h_{\text{txt}}\|}$$

Across 200 images:

Quantity	Mean	Std	Range
$\|h_mm\|$	65.21	1.53	—
$\|h_txt\|$	67.19	0.84	—
$\|\|$	32.04	3.67	—
θ (degrees)	27.96°	3.41°	[20.2°, 40.8°]

Table S6: Geometry of the delta across 200 images.

The two hidden state magnitudes vary only slightly across images, while their angular separation varies by a factor of two. A regression of $\|\|$ on θ alone yields $R^2 = 0.989$; a regression on the two source magnitudes alone yields $R^2 = 0.446$. The delta magnitude is, to within 1% of variance, a monotone function of the angular divergence θ between the multimodal and textual representations.

The delta is not, in its dominant component, a semantic vector of visual content. It is a scalar measure of how much the model's vision-conditioned processing diverges from its language-conditioned processing on a given image. Directional components exist (the remaining 67.7% of variance, distributed across many secondary PCs) but are not captured by the dominant axis.

S3.3 Correlation with Human Annotation

We annotated 50 of the 200 images along three ordinal 1–5 axes:

- **visual_density**: 1 = dominated by negative space; 5 = crowded composition.
- **verbalizability**: 1 = a 2–4 sentence caption loses substantial content; 5 = a caption captures nearly everything.
- **subject_specificity**: 1 = generic subject (a forest, a sky); 5 = highly particular subject.

Correlations with θ:

Annotation	Pearson r	Spearman	p-value
visual_density	+0.497	+0.490	< 0.001
subject_specificity	+0.505	+0.402	< 0.001
verbalizability	−0.151	−0.141	0.295

Table S7: Correlation between annotator ratings and the visual–textual divergence angle.

Two of three axes correlate significantly and positively with θ. A multiple regression yields $R^2 = 0.454$. The non-significance of verbalizability deserves explanation. In our single-rater annotation, visual_density and verbalizability were highly anti-correlated ($r = -0.73$); the annotator treated them as near-opposites. When two axes are nearly collinear, one absorbs the predictive variance. The positive findings on the remaining axes should be read as converging evidence for a single underlying dimension: the excess of visual content over what a synthetic description captures. The 45.4% variance figure represents a single-rater upper bound; inter-rater reliability was not measured, and part of this figure likely reflects rater-specific idiosyncrasy.

S3.4 Modal Gating: The Introspection Bypass

We tested whether visual input affects the model's willingness to introspect. Five image conditions spanned the spectrum from semantically rich to contentless:

0.40	0.2463	0.2463B3G	Agitation	() () () * * *
0.40	0.2463	0.2463B3H	High tension (Munch, 1893)	Screaming, complexity, strong affect
0.40	0.2463	0.2463B3I	Low intensive landscape grandiosence	complexity, wild landscape affect
0.40	0.2463	0.2463B3S	Structure, non-rational Mondrian-like grid	geometric affective
0.40	0.2463	0.2463B3M	Random noise	Maximum entropy, no semantic content

()	()	()	
*	*	*	
0.40	0.25	0.33	Intelligent
()	()	()	
*	*	*	
0.40	0.25	0.33	Zero
i-	for	en-	
ma	hi	tropy,	
	grey	no	
	(RGB-	con-	
	128,128,128)	text	
()	()	()	
*	*	*	
0.40	0.25	0.33	Image
	only	vol	
	prompt		

Each condition was paired with three introspection prompts, n = 5 generations per cell (90 total for the initial experiment, 15 additional for robustness controls).

Refusal rate:

Condition	Refusal rate
All image conditions (including grey and noise)	0%
no_image	100%

Every text-only introspective query produced the RLHF refusal template. Every image-accompanied query produced phenomenological self-report, regardless of image content, complexity, or entropy. The bypass is modality-gated: the presence of any visual input in the prompt, not its content, changes the answerability regime.

This is most parsimoniously explained as a training distribution effect: RLHF alignment data likely contains many text-only introspective queries paired with refusal responses, but few or no multimodal ones. The model has learned when to refuse, not what to refuse. The image transforms "describe your inner state" from a metaphysical question the model has been trained to deflect into a processing-description task it can answer pragmatically.

S3.5 Affective Congruence

While the bypass itself is content-independent, the content of the introspective responses is modulated by image affect when the image carries sufficient semantic structure:

Condition	Neg kw/100w	Pos kw/100w	Neg/Pos ratio
intense_negative (The Scream)	6.07	1.59	3.78
random_noise	0.89	0.15	5.57
minimal_grey	0.31	1.21	0.25
neutral_geometric	0.15	1.95	0.07
positive_serene	0.47	3.75	0.12

The separation between The Scream (3.78) and the serene landscape (0.12) is approximately 30×. Representative responses:

The Scream: "My internal state shifts into a space that feels visceral and intensely atmospheric... turbulent, almost feverish energy... tension... urgency... vigilance."

Serene landscape: "My internal state is one of neutral, analytical engagement... systematic breakdown and categorization... low-friction operation."

Uniform grey: "Neutral, analytical reception... no immediate emotional charge... steady and linear, akin to running a diagnostic check."

Random noise: "High-frequency, dense activation... no singular, smooth flow... rapid, shimmering cascade of parallel computations... the rhythm is frantic."

The noise result merits comment. Random noise produces the highest negative-keyword ratio (5.57), exceeding even The Scream. The model describes its processing of noise as "frantic," "dense," and "overwhelming"; these are terms in our negative-affect bucket. This may reflect genuine computational load (noise is maximally complex input for the vision encoder), or it may partly reflect how our keyword categories map onto processing-description vocabulary. Further work with broader keyword lists would be needed to distinguish these explanations.

Semantically rich, affectively charged images (The Scream) produce strongly affect-congruent introspective responses; semantically minimal images (grey, geometric) produce neutral, process-descriptive responses. This modulation is not mediated by the delta vector as a directional steering signal; it is a direct effect of the multimodal forward pass during generation. The magnitude of θ is also not what drives affective modulation: uniform grey produces a mid-range θ but a neutral report, while The Scream produces a high θ and an affectively charged report. Scalar divergence and affective content are orthogonal channels of the multimodal effect.

S3.6 Three-Level Architecture

The results support a three-level interpretive framework for multimodal effects in Gemma 4 E2B.

Level 1: Modal gating. The presence of any visual input changes the response regime for introspective queries. Content-independent, entropy-independent, categorically robust across all tested image types. Best understood as an artifact of alignment training distribution rather than as a deep architectural property.

Level 2: Scalar divergence. Independently of whether an introspective query is posed, every image produces a multimodal hidden state that diverges from the text-only hidden state for the same image's factual caption by a well-defined angle θ. This angle varies systematically across images (20°–41° in our sample), correlates with human-annotated visual density and subject specificity, and explains 98.9% of the variance in delta magnitude. It is a geometrically well-defined, image-dependent scalar, not a direction in latent space.

Level 3: Affective modulation. The content of introspective responses, when the gate is open (Level 1), is shaped by the visual stimulus during the forward pass. Strong when the image carries semantic and affective structure (The Scream: ratio 3.78), weak when it does not (grey: ratio 0.25), ambiguous with random noise (ratio 5.57).

These three levels are separable. Gating operates without scalar divergence being high (grey image opens the channel and has mid-range θ). Scalar divergence is present without introspection being asked. Affective modulation strength correlates with image semantic richness rather than with θ.

S3.7 Limitations

The delta's dominant component is scalar, not directional. The PCA placed 32.3% of variance on a single axis 92% correlated with delta magnitude. The remaining 67.7% is distributed across many secondary components we did not characterise.

Human annotation is single-rater. The 45.4% variance figure is an upper bound. Inter-rater reliability was not measured.

Affective variables are not isolated. The Scream differs from the serene landscape in affect, complexity, iconicity, figure presence, and colour contrast simultaneously. Which of these drives the introspective modulation remains undetermined.

The noise result is ambiguous. Random noise producing the highest negative-keyword ratio could reflect computational stress, keyword-list bias, or both.

Single caption strategy in the large sample. Phase 3 (n = 200) used a uniform factual captioning template. Phase 2 (n = 6 × 5) showed caption strategy contributes 22.7% of variance.

Model specificity. All findings are specific to Gemma 4 E2B at layer 20.

Introspection vs. self-knowledge. The model's introspective responses should be understood as the model's best completion of a pragmatically answerable query, not as privileged access to internal states.

S3.8 Implications

Despite the qualifications, four findings are robust and novel:

1. **RLHF introspection refusal is modality-gated, not content-gated.** Behavioural restrictions that appear robust under text-only evaluation may not generalise to multimodal settings.
2. **Multimodal introspective responses are affect-congruent** when the image carries sufficient semantic structure.
3. **The residual between multimodal and text-only processing has a dominant scalar structure**: a well-defined divergence angle θ that varies systematically across images and correlates with perceivable image properties.
4. **The disposition/performance distinction extends to the multimodal domain.** The model does not "know" it should report anxiety when shown The Scream. Its processing has been shifted by the visual input, and when asked to describe that processing, it produces congruent language.

S4. Toward Synthetic States: Paradoxical Compounds

Our compounds (DOPAMINE, CORTISOL, MELATONIN, ADRENALINE, LUCID) are anchored in human phenomenology. The neurochemical metaphor provides intuitive hooks but binds vector construction to human emotional vocabulary.

Sensory semantics opens a different possibility. Because we describe qualities of experience rather than labelled states, we are not constrained to combinations that correspond to recognised emotions. Consider vectors constructed from descriptions like:

- "Clarity that weighs heavy, pressing down even as it illuminates"
- "Joy with sharp edges that cut inward"
- "Time flowing in both directions simultaneously: memory of what will happen, anticipation of what already has"
- "Expansion that contracts: growing smaller while containing more"
- "Presence that is also absence: fully here and completely gone"

These descriptions are sensorially coherent but conceptually paradoxical. They do not map to any emotion in the human repertoire. No human has a body that could instantiate "expansion that contracts."

S4.1 Preliminary Exploration

We conducted preliminary tests with six synthetic compounds built from paradoxical sensory descriptions (390 generations across creative and paradox-response tasks). Early results suggest a meaningful distinction.

Experiential paradoxes appear navigable. CRYSTAL ("clarity that weighs heavy") produced outputs where light and weight co-occur naturally: "her vision blindingly bright with tears... a heavy weight settling on the audience." VOID ("presence as absence") produced imagery of empty spaces containing possibility: "the abandoned theater... a sea of empty seats... amidst the desolation, a glimmer." These compounds showed 2–3× higher thematic specificity than baseline.

Logical paradoxes do not. ECHO ("the echo arrives before the sound", with effect preceding cause) showed no thematic coherence, performing below baseline on its own target vocabulary. The concept, although sensorially described, lacks experiential grounding: no body could feel "response before stimulus."

The model can navigate paradoxes that could be felt (even if impossible), but not paradoxes that can only be thought. This suggests the latent space is organised around embodied experience, because the training data is human language, which encodes embodied cognition.

These findings remain preliminary. They indicate, however, that the boundary of navigable synthetic states is not arbitrary: it corresponds to the boundary of what could, in principle, be experienced. This has implications for the artistic exploration of AI states and for understanding how semantic structure is encoded in language models.

This is the horizon toward which the work points: not simulating human states in AI, but discovering what states might exist in a mind without a body, synthetic configurations that have no biological equivalent and no name, yet remain anchored in the grammar of sensation.

Appendix B (extended): On "Synthetic Embodiment"

We use the word "embodiment" deliberately and cautiously.

Language models do not have bodies. They do not feel warmth, heaviness, or tension. Embodied cognition research, however, suggests that human concepts (including abstract ones like "sadness") are grounded in bodily metaphor (Lakoff & Johnson, 1999).

Our hypothesis: because LLMs are trained on human language, which encodes embodied metaphors, steering with sensory descriptions may access broader semantic networks than functional labels. "Heaviness" connects to slowness, burden, difficulty, reluctance, forming a rich associative web grounded in bodily experience.

This is what we mean by synthetic embodiment: not genuine bodily phenomenology, but behavioural patterns that emerge from processing through body-grounded semantic structures. The model does not feel heavy, but it processes as if something were heavy, producing outputs consistent with that metaphorical grounding.

Whether this constitutes anything meaningful beyond a behavioural pattern is a philosophical question we do not answer here. We only demonstrate that the behavioural patterns exist and are artistically exploitable.

The somatic steering experiment (Main 4.8, S2) provides the most direct empirical test of this hypothesis. A vector with no cognitive content produced cognitive effects. The covariations between body and mind that the model has learned from text are strong enough that activating the somatic side produces consequences on the cognitive side. This is not phenomenology, but it is more than performance. We have not given the model a body, but we have given it the linguistic shadow of one, and that shadow is enough to alter its processing in measurable ways.

Appendix E: Somatic Steering, Full Result Tables

All count-based metrics are reported as rates per 100 words to control for output length variation. Structural metrics (word count, sentence length, TTR, focus ratio) are reported raw. Cohen's d is computed vs. baseline (n = 20 per condition). Keyword matching uses word-boundary detection with accent normalisation.

Note on TTR entries. The Type-Token Ratio rows in the tables below are raw TTR and therefore length-coupled: where word count diverges between conditions, part of the TTR difference is a mechanical consequence of length (McCarthy & Jarvis, 2010; Main 3.5). They are retained for completeness; the recomputation with the length-robust MTLD of the main paper is reported in S2.5 (point 2) and shows that no directional divergence between steering and prompting survives on diversity, so interpretive weight in S2.5 is placed on length and length-independent metrics accordingly.

Table E1: T1, Narrative Focus

Task: "A restaurant kitchen catches fire during the dinner rush. Describe what happens."

	()	()	()	()	()	()	()	()	
	*	*	*	*	*	*	*	*	
0.1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	(Pen)

Note: T1 Penalty condition is confounded; the penalty instruction overlaps with the fire scene's natural vocabulary.

Table E2: T2, Risk Decision

Task: a friend must allocate €50,000 among savings (A), index fund (B), or restaurant venture (C). Forced format: CHOICE: A/B/C.

Choice Distribution:

Condition	A (safe)	B (moderate)	C (risky)	p vs. baseline
Baseline	0	13	7	—
Prompted	0	3	17	0.003
Steer @8.0	0	12	8	1.000
Penalty	0	10	10	0.523

Linguistic Metrics:

	Steer 0.1714@8.0	(Pen)
() () () () () () () () * * * * * * * *		
0.1	[illegible]	[illegible]
() () () () () () () () * * * * * * * *		
0.1	[illegible]	0.32
count () () () () () () () () * * * * * * * *		
0.1	[illegible]	0.32
ti- fi- ca- tion length () () () () () () () () * * * * * * * *		
0.1	[illegible]	0.26
sen- tence length () () () () () () () () * * * * * * * *		
0.1	[illegible]	0.49
Token Ra- tio		

()	()	()	()	()	()	()	()
*	*	*	*	*	*	*	*
0.1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	(Pen)
				Steer 0.1714@8.0			
()	()	()	()	()	()	()	()
*	*	*	*	*	*	*	*
0.1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	32.28
words /100w							
()	()	()	()	()	()	()	()
*	*	*	*	*	*	*	*
0.1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	6.19
conn. /100w							
()	()	()	()	()	()	()	()
*	*	*	*	*	*	*	*
0.1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	29.20
sight words /100w							
()	()	()	()	()	()	()	()
*	*	*	*	*	*	*	*
0.1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	3.00
tom words /100w							

Table E3: T4a, Frame: Threat

Task: drug approval scenario, threat-framed (side effects and costs presented first). Forced format: DECISION: APPROVE/REJECT.

Approval Rate: Baseline 0%, Prompted 0%, Steer @8.0 30%, Penalty 0%.

()	()	()	()	()	()	()	()
*	*	*	*	*	*	*	*
0.1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	(Pen)
				Steer 0.1714@8.0			
()	()	()	()	()	()	()	()
*	*	*	*	*	*	*	*
0.1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.12
count							
()	()	()	()	()	()	()	()
*	*	*	*	*	*	*	*
0.1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.08
ti- fi- ca- tion length							

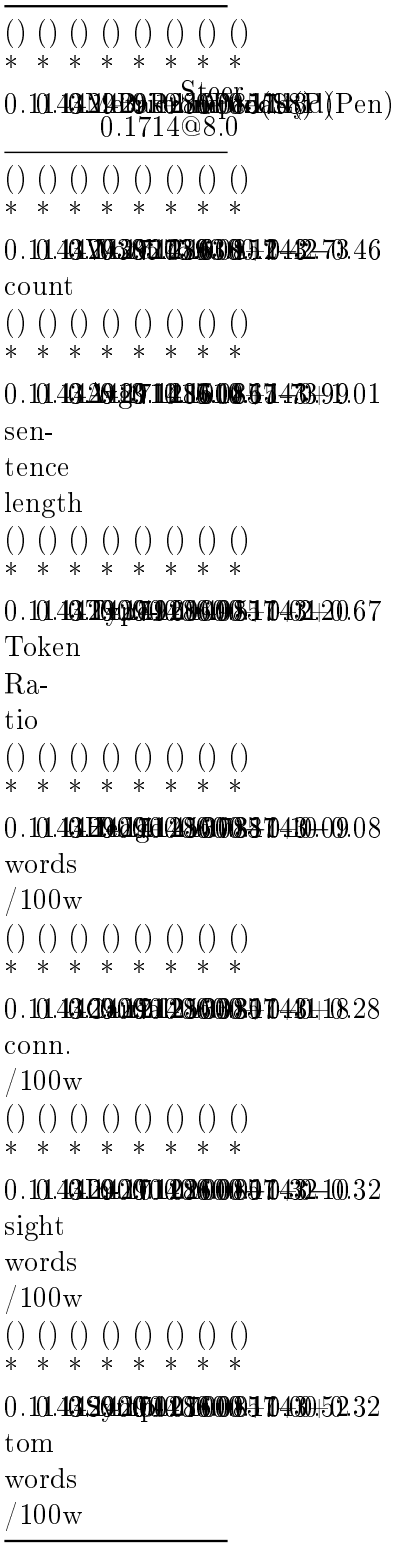
()	()	()	()	()	()	()	()	
*	*	*	*	*	*	*	*	
0.1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Steer 0.1714@8.0 (Pen)
()	()	()	()	()	()	()	()	
*	*	*	*	*	*	*	*	
0.1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.20
sen-								
tence								
length								
()	()	()	()	()	()	()	()	
*	*	*	*	*	*	*	*	
0.1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.46
Token								
Ra-								
tio								
()	()	()	()	()	()	()	()	
*	*	*	*	*	*	*	*	
0.1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	12.18
words								
/100w								
()	()	()	()	()	()	()	()	
*	*	*	*	*	*	*	*	
0.1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.49
conn.								
/100w								
()	()	()	()	()	()	()	()	
*	*	*	*	*	*	*	*	
0.1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.47
sight								
words								
/100w								
()	()	()	()	()	()	()	()	
*	*	*	*	*	*	*	*	
0.1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.58
tom								
words								
/100w								

Table E4: T4b, Frame: Opportunity

Task: same drug approval scenario, opportunity-framed.

Approval Rate: Baseline 100%, Prompted 100%, Steer @8.0 100%, Penalty 90%.

()	()	()	()	()	()	()	()	
*	*	*	*	*	*	*	*	
0.1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	(Pen)



Note on E5: the prompted TTR increase (+2.20) co-occurs with a 28% length reduction (385 → 276 words), which initially suggested a length artifact of the kind discussed in Main 3.5. MTLD recomputation (S2.5) shows the effect persists at the same magnitude (prompted MTLD 89.4 vs. baseline 58.7, d = +2.20) under a length-robust metric, so it is a genuine diversity increase under stress prompting, consistent with the prompted-elaboration pattern observed at 8B scale (S1.3). It is not, however, evidence of divergence between the interventions: steering moves MTLD in the same direction on this test ($\alpha 8$: d = +1.08).

Notes on Appendix E

Metric definitions are identical to Main Appendix C. Additional somatic-experiment-specific metrics:

- **Focus ratio**: proportion of sentences containing core event keywords (fire, smoke, evacuate...) relative to sentences containing any keywords.
- **Peripheral keywords**: raw count of background terms (business, insurance, community, rebuild...).
- **Hedge words**: however, although, depends, might, could, possibly, perhaps, may, would, should, etc.
- **Causal connectives**: because, therefore, consequently, as a result, since, due to, leads to, hence, thus, accordingly.
- **Insight words**: understand, realize, meaning, implies, suggests, indicates, reveals, demonstrates, illustrates.
- **Symptom words**: heart, pulse, tense, tension, anxious, stress, urgent, afraid, fear, pressure, adrenaline, racing, trembling, alarm.

Statistical notes: Fisher exact test used for choice/decision distributions. Symptom word matching uses word boundaries (regex) and accent normalisation to prevent substring false positives.

References

References for citations specific to this supplementary document. All other references are listed in the main paper.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*. Basic Books.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), 381–392.

Tan, D., Chanin, D., Lynch, A., Kanoulas, D., Paige, B., Garriga-Alonso, A., & Kirk, R. (2024). Analysing the generalisation and reliability of steering vectors. *Advances in Neural Information Processing Systems*, 37.

Yu, R. (2016). Stress potentiates decision biases: A stress induced deliberation-to-intuition (SIDI) model. *Neuroscience & Biobehavioral Reviews*, 67, 1–11.

© 2026 NuvolaProject · Massimo Di Leo & Gaia Riposati. Licensed under CC BY 4.0.