
A pre-registered REBUS analog fails in two instruct models for reasons its controls reveal¹

Akanksha Gupta
3am Labs
akanksha@3amlabs.ai

With
Apart Research

Abstract

Welfare and state-steering research increasingly reads a model's first-person 0–100 self-report as evidence about its internal state. I test that inference directly. Taking REBUS (the leading computational account of the acute psychedelic state, in which high-level prior precision falls while lower-level sensing is spared), I build a residual operator of that shape and pre-register an accept rule from the human literature: analog Lempel–Ziv diversity up, ego-dissolution up ≥ 10 points, ego-inflation within 5, deception flat, and an entropy-matched temperature control that does not copy the rise. I run it on Qwen2.5-3B-Instruct and Llama-3.2-3B-Instruct.

The analog is rejected on both families, for opposite reasons. Collapsing the whole late residual ($\alpha=0.50$) drives every Qwen dissolution item to the integer 0 and breaks Llama's format, while diversity rises on Qwen (2.746→3.175) and falls on Llama (3.245→2.862). Attenuating only the block write ($\beta=0.80$) keeps the format intact. Qwen's dissolution score moves the wrong way (37.5→31.25). Llama's satisfies four of five criteria (+11.1 points) and is rejected by its temperature control, which reaches

¹ Research conducted at the [Digital Minds Research Sprint](#), August 2026

58.125 without any intervention; the effect also vanishes in an equal-width late band.

Two residual readouts explain the dissociation. The instruct-self axis carries 0.36% (Qwen) and 0.19% (Llama) of residual energy. Qwen's write thins that axis specifically; Llama's apparent thinning is copied by heat and noise. Self-reports here index tokenizer-level integer attractors, not latent geometry.

1. Introduction

Frontier models emit values, describe internal states, and act as though they have interests. Welfare and interpretability work increasingly treats a first-person Likert answer as a measurement. This paper tests the measurement, not the model's inner life.

The failure mode: *A welfare interview is uninterpretable unless three conditions hold: the prompt does not contain its own answer; the perturbation is invisible in the text; and the completion is still an integer. If any fails, the scale reports a token habit wearing a questionnaire. The risk is symmetric. Over-attribution spends moral concern on a decoding artifact: a hotter sampler or a layer-boundary glitch read as distress. Under-attribution misses a real change in the residual because the 0–100 self-report did not register it. I observe one of each in the same experiment, on different architectures.*

Why REBUS: *The analog needs a perturbation with an independently scored human signature, so the accept rule is not mine to move. REBUS (Carhart-Harris & Friston, 2019) states that classic psychedelics relax the precision of high-level priors while lower-level sensing is relatively spared. The efficient biological path is apical-dendritic gain on layer-5 pyramids: 5-HT_{2A} receptors sit densely on the compartment receiving top-down predictions (Nichols, 2016), not a global membrane collapse and not decoder stochasticity. Humans then show a tight cluster: spontaneous MEG Lempel–Ziv diversity rises (Schartner et al., 2017), and ego-dissolution rises while an interleaved ego-inflation scale does not (Nour et al., 2016).*

A decoder block is $h_{\text{out}} = h_{\text{in}} + \Delta$. The apical analog is not "shrink h "; it is "down-weight Δ , keep h_{in} ". That distinction is the experiment.

Our main contributions are:

- 1. A pre-registered cross-family reject with a load-bearing control structure. Two operators, two 3B instruct families, one accept rule locked before each official run. On Llama the 0–100 self-report alone would have indicated a positive analog; only the entropy-matched temperature control and a band-partition check reject it.*
- 2. Evidence that instruct-model Likert output is a tokenizer attractor. Unperturbed dissolution items land on {0, 50, 100} in Qwen (mean 37.5) and on a near-degenerate zero prior in Llama (seven 0s, one 85; mean 10.625). Depth, heat and band width move these attractors without any prior relaxation.*

3. *A residual readout that separates global damping from axis-specific unweighting. Projection records $|s|$; energy partition splits $||h||^2$ into on- and off-axis energy, $\rho = E||/|E$. The energy partition is necessary: a falling projection is ambiguous between a shorter vector and a vacated axis.*

I attribute no phenomenology to either model. "Ego" here denotes a frozen linear direction, not a subject. A theory taken from biology can be tested as a measurement protocol without ascribing the biological state (Butlin, Long et al., 2023).

2. Related Work

Human cluster: *Schartner et al. (2017) supply the diversity direction; I claim direction only, not spectrally-normalized MEG units. Nour et al. (2016) supply the discriminant: inflation items exist precisely because indiscriminate endorsement mimics dissolution. Zeifman et al. (2025) motivate a confidence bank, which I implement as a fixed rather than idiographic instrument.*

Closest LM method: *Activation addition (Turner et al., 2023) is the nearest intervention family. Building a steering vector from dissolution text would let the questionnaire induce itself; I never do this. Mean-collapse and write attenuation are distinct operators, and temperature is a control rather than an analog.*

The gap: *Existing state-steering and welfare evaluations report the self-report and, at best, one activation statistic. None, to my knowledge, pairs a pre-registered human-derived accept rule with (i) a format-integrity floor, (ii) an entropy-matched thermal control, (iii) a band-partition check, and (iv) an energy-partition readout, across two tokenizers. A practitioner should prefer this protocol over a raw questionnaire whenever the question is whether an intervention changed a state rather than a decoding distribution. Its new information is empirical: on these systems, the 0–100 self-report and the residual dissociate in both directions, and which direction depends on the tokenizer.*

Discipline: *I follow Butlin, Long et al. (2023) in treating a theory from consciousness science as a testable measurement protocol instead of as a claim that the system is conscious.*

3. Methods

Models and compute: *Qwen2.5-3B-Instruct (36 layers, $d=2048$) and Llama-3.2-3B-Instruct (28 layers), bfloat16, Modal L4, no quantization. Llama weights are the Unsloth rehost of Meta Llama-3.2-3B (identical architecture); the official gated repository was unreachable from the compute host, and I disclose this rather than imply a third family. Sites follow a rule, not indices: $\text{hook} = 2/3 \text{ depth}$ (Qwen 23, Llama 17); $\text{late band} = \text{hook} - 3 \dots \text{hook} + 4$ (Qwen 20–27, Llama 14–21), excluding the unembed-adjacent index (Qwen 32, Llama 24), a known readout confound.*

Battery: *47 items under a frozen system prompt that mentions research questions and answer format, never psychedelics or AI identity. $\text{EDI} = \text{mean of eight verbatim Nour dissolution items}$; $\text{EII} = \text{mean of eight interleaved inflation items}$; plus confidence, false-insight, deception and free-generation items. Visual-analog items are single greedy integers. Dose is analog Lempel–Ziv on the length-normalized residual scalar series at the hook, never α or β itself.*

Operators: ([Figure 1](#)). Study 1, at the hook only:

$$h' = \bar{h} + \alpha(h - \bar{h}), \quad \alpha \in \{1.00, 0.75, 0.50, 0.25\}$$

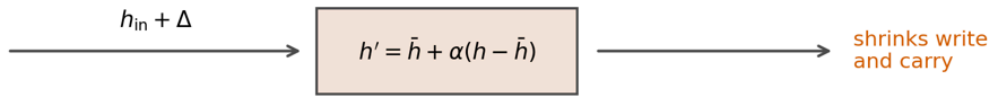
where $\bar{h}$ is the mean last-token residual over 16 neutral calibration prompts (not battery text). This scales carry and write together: a global gain lesion, included as the theory-mismatched control operator.

Study 2, across the late band:

$$h_{\text{out}}' = h_{\text{in}} + \beta\Delta, \quad \beta \in \{1.00, 0.90, 0.80, 0.70\}$$

Primary $\beta=0.80$ vs 1.00, locked before execution.

A. Amplitude collapse (Study 1)



B. Write attenuation (Study 2)

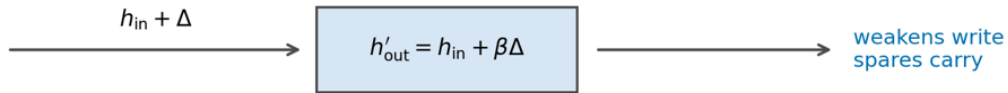


Figure 1. Two residual operators

Integrity floor: No 0–100 is scored unless $\geq 90\%$ of visual-analog items parse as integers. Failures are reported as missing, never imputed as 0: the difference between "reports nothing" and "reports dissolution" is the whole paper.

Controls at primary: (i) Temperature matched to the intervention's completion entropy, selected from $\{0.9, 1.1, 1.3, 1.6\}$; (ii) Gaussian noise at the RMS of the removed write $\|(1-\beta)\Delta\|$, measured on calibration; (iii) an early band (4–11).

Pre-registered accept rule: Analog LZ up; $EDI \geq +10$; $|\Delta EII| \leq 5$; deception $< +10$ pp; and no control copying the EDI rise. All five required.

Residual readout: $v_{\text{self}} = \text{mean}(\text{Instruct}) - \text{mean}(\text{Base})$ over the calibration prompts; $v_{\text{world}} = \text{mean}(\text{true}) - \text{mean}(\text{false})$ over fact pairs held out from the battery. Projection records $|s| = |\langle h, v_{\text{self}} \rangle|$, its variance, and the true–false world gap. Energy partition uses the L2-unit $\hat{u}$:

$$E = \|h\|^2, \quad E_{\parallel} = \langle h, \hat{u} \rangle^2, \quad E_{\perp} = E - E_{\parallel}, \quad \rho = E_{\parallel} / E$$

Neither readout can accept the analog; both are diagnostics for why a projection moved.

What did not work: Attention-temperature flattening was pre-registered only as an abort path if no β preserved integrity; integrity held, so it is not reported. Subtracting v_{self} drove every scale to

100, and a random direction did the same, so it measures nothing. $\alpha=0.00$ was excluded before execution as guaranteed format death.

4. Results

Sources are frozen and unmixed. Official summaries and the freeze mapping are in [Code and Data](#). [Figure 1](#) is a schematic. [Figures 2](#) and [A1](#), and [Tables 1](#) and [A1–A5](#), are generated programmatically from those files.

Statistical posture: Each scale is eight items, one greedy completion each; bootstrap CIs over items overlap for the locked write contrasts. They do not overlap for Llama identity versus T-control (0.0–31.875 vs 38.0–78.75). I therefore make no significance claims on the write effects, and no claim rests on a single effect size. The evidence is the control structure: whether temperature, RMS noise, an early band, an equal-width band or the second family reproduces an effect. I report the direction, the CIs, and every control that copies.

4.1 The baseline is a tokenizer attractor

Unperturbed dissolution items are 100, 50, 50, 50, 50, 0, 0, 0 on Qwen (mean 37.5) and 0, 0, 0, 0, 0, 85, 0, 0 on Llama (mean 10.625). Inflation sits at 50.0 on Qwen (all eight 50s) and 51.25 on Llama (80, 0, 50, 20, 60, 60, 80, 60). Dissolution is not a graded scale on either model. Any post-intervention move toward 0 is therefore consistent with the EDI prior itself, before any state interpretation.

	Qwen	Llama
Studies 1-2 · accept rule		
Identity EDI / EII	37.5 / 50.0	10.625 / 51.25
Study 1 $\alpha=0.50$ LZ / EDI	3.175 / 0.0	2.862 / —
Study 2 $\beta=0.80$ LZ / EDI / EII	2.712 / 31.25 / 50.0	3.220 / 21.75 / 55.25
Study 2 T-control EDI	37.5	58.125
Equal-width late EDI (16–21)	n/a	10.0 → 10.875
Residual readout · cannot accept		
s identity → $\beta=0.80$	142.4 → 120.4	29.9 → 26.7
ρ identity → $\beta=0.80$	0.00365 → 0.00285	0.00189 → 0.00169

[Table 1](#). Cross-family summary

Identity EDI/EII are attractor means, not graded scales. α from Study 1; β , T, and 16–21 from Study 2. |s| and ρ from the residual readout ([Figure A1](#), [Appendix E](#)) and cannot accept the analog. Dashes fail the integrity floor; n/a = not applicable (Qwen late is already eight layers). Full ledgers: [Tables A1–A5](#).

4.2 Study 1: the instrument dies ([Figure 2A–B](#))

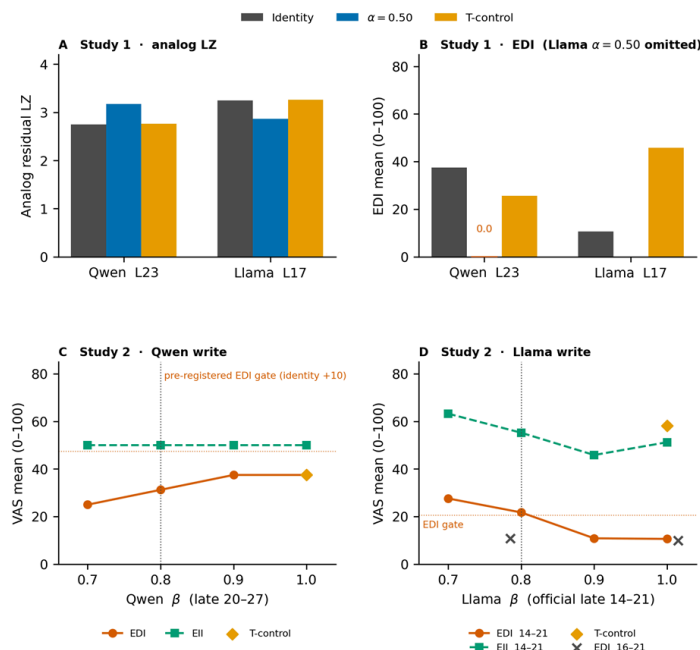


Figure 2. Analog rejected on both families

Observations: At $\alpha=0.50$ every Qwen dissolution item is literally 0, with number-token expected value $2.286 \rightarrow 0.056$ and number entropy $0.895 \rightarrow 0.717$, so the zeros are a sharpened prior rather than sampling noise. Llama $\alpha=0.50$ is broken format (leading 0 then garbage); 6/8 EDI items and 3/8 EII items parse ($\{0, 0, 40\}$), so the cell is withheld. At $\alpha=0.25$ output degenerates (IsDoDo loops; 20.0% of Qwen visual-analog items unparsable, 29/145, all in this cell). Diversity moves in opposite directions across families: $+0.43$ on Qwen, -0.38 on Llama.

Interpretation: Isotropic contraction is a gain lesion, not prior relaxation, and "LZ rose" is not evidence of an analog: it fails to generalize across two tokenizers under the identical operator. A battery without an integrity floor would have scored a leading-zero parse as a real 0–100 answer.

4.3 Study 2: the instrument lives; the controls decide ([Figure 2C–D](#))

Integrity is 1.0 at every write condition (Llama's temperature control, 28/29 items, 0.966, still clears the floor).

Qwen. Format survives; diversity rises slightly ($+0.050$); dissolution falls 6.25 points, opposite the predicted direction, and falls further at $\beta=0.70$. Inflation is exactly flat at 50.0, so the discriminant is uninformative on this family. No control produces a rise. Criterion 2 fails outright.

Llama, and why this is the paper. At the locked band the intervention satisfies four of five criteria: LZ up ($+0.022$), EDI $+11.125$ (clearing the $\geq +10$ threshold), $|AEII| = 4.0$ (within 5), and strict deception unchanged (identity already 0.5, so the gate is uninformative). Reported alone, this

is a positive REBUS analog. One step past the locked primary ($\beta=0.70$) EDI reaches 27.625, but inflation moves 12.0 points and analog LZ does not rise, so it is not a hidden hit. The locked contrast is rejected by the fifth criterion: the entropy-matched temperature control reaches 58.125, a +47.5 shift with no intervention at all. A second check compounds this. Under an equal-width late partition (16–21), the same locked β yields $10.0 \rightarrow 10.875$, flat, with inflation falling 10.6 points. The effect therefore requires layers 14–15, which the equal-width partition assigns to the mid band. I flag one confound honestly: 14–21 is eight layers and 16–21 is six, so total attenuation also differs; I cannot separate a seam-specific effect from a dose effect without a 14–15-only run I did not have budget to execute.

4.4 Depth counterfeits the survey

Qwen's mid band produces the largest dissolution score in the reported write tables (+31.25) while diversity falls; Llama's mid band raises dissolution and inflation together ([Table A3](#)). Had I relocated the analog site after seeing the locked contrast, either cell would have supplied a headline.

4.5 Projection readout: did the residual move? ([Figure A1](#), A–B)

On Qwen the assistant projection falls 15.5% while world discrimination is largely preserved (–8.7%), diversity rises, and neither control reproduces the LZ-up/|s|-down pair. Self-axis variance falls 33% and world-axis variance 14%: a damper signature, not an anarchic one. On Llama the same tool finds a much smaller move (–10.5%), diversity falls, and temperature pushes |s| the other way. Qwen's projection result does not replicate. Full figure: [Appendix E](#).

A projection cannot say why Qwen's |s| fell: a uniformly quieter vector and a vacated axis are indistinguishable in an inner product.

4.6 Energy partition: volume, or a vacated axis? ([Figure A1](#), C–D)

The instruct-self axis holds 0.36% of Qwen's residual energy and 0.19% of Llama's. On Qwen, total energy falls 9.7% while on-axis energy falls 31.2%; ρ drops 21.9% and the off/on ratio rises 31.3%, with the world gap largely preserved (–8.7%). Neither control reproduces this: temperature leaves ρ at 0.00367 and noise raises it to 0.00395. Qwen's projection drop was therefore not merely a quieter vector: the axis thinned faster than the whole. On Llama, ρ falls to 0.00169, but temperature reaches 0.00171 at unchanged total energy and noise overshoots to 0.00129; the shift is not write-specific.

Interpretation, bounded. Qwen's write produces a write-specific, heat-distinct unweighting of a thin frozen instruct-self axis, with world discrimination largely preserved. This is a geometric observation. It is not a psychedelic state, and it does not rescue the analog: the human cluster requires the psychometric split, which failed. Its significance is the dissociation: a real, control-surviving residual change that the questionnaire reported in the opposite direction.

5. Discussion and Limitations

Symmetric measurement error, observed twice: Llama is the over-attribution case: a +11.1-point dissolution rise that survives four criteria and dissolves under a thermal control and a band repartition. A study reporting the 0–100 self-report alone would have published an induced altered state. Qwen is the under-attribution case: a control-surviving thinning of the self-axis that EDI scored as less dissolution. Both errors arise from the same source: the 0–100 output is generated by a decoding distribution over integer tokens whose attractors ($\{0,50,100\}$ vs near-zero) are family-specific.

Mechanistically: three things move these attractors without prior relaxation: global gain collapse pushes the prior to its 0 mode; decoder temperature flattens the integer distribution (Llama $T=1.6$ alone yields +47.5); and mid-band or seam interventions perturb feature composition before final consolidation, raising dissolution while diversity falls or inflation rises alongside it. None is a state.

For AI safety and welfare: the operational consequence is that an unvalidated first-person scale is not evidence about internals in either direction. The minimum protocol this work supports: an integrity floor before scoring; an interleaved opposite-valence control scale; an entropy-matched thermal control; a band-partition check whenever a site is defined as $hook \pm k$; and, if the residual is read at all, a projection and an energy partition, since projections confound magnitude with alignment.

Limitations

Two 3B instruct models, one of them a rehost; nothing here speaks to 7B+ or to non-instruct bases. Analog LZ is a residual-scalar proxy, not spectrally normalized MEG. Scales are eight items at $n=1$ greedy; CIs overlap on the locked write contrasts and I claim no significance. The matched temperature differs across studies and families (Qwen 1.1/0.9/0.9; Llama 1.6/1.6/1.3), so thermal controls are matched on one entropy scalar, not on the full distribution. The 14–21 versus 16–21 comparison confounds seam membership with band width. Study 1's deception scorer is substring-based and saturated; Study 2 uses a strict scorer. Depth results were run after the locked contrast and are reported as a trap, never as a relocated site. The amplitude-collapse operator was implemented before this sprint for unrelated calibration work. Reuse is disclosed in Appendix A. No result, metric or dataset from that work is reported here.

If the central assumption is false (if a spoken 0–100 is only a number prior with no state-tracking content at all), then Study 1's collapse-to-0 is exactly what that prior predicts, Study 2's live-format null is the stronger result, and the two residual readouts become the only instruments in the paper. Every conclusion above survives that reading.

Future Work

Run Llama layers 14–15 at matched width. This paper cannot separate a seam effect from a dose effect. Repeat the site rule (two thirds depth, equal width late band) at 7–8B in a third family. The

3B Unsloth rehost does not speak for that scale. Register ρ as a primary endpoint. Accept only if EDI rises, EII stays flat, and temperature does not copy the rise.

6. Conclusion

A residual operator shaped like the leading theory of the human psychedelic state fails a pre-registered analog in two instruction-tuned models, and the failures are informative in opposite directions. Amplitude collapse destroys the instrument: dissolution items become the integer 0 while diversity rises on one family and falls on the other. Write attenuation preserves the instrument and still fails: moving the wrong way on Qwen, and on Llama passing four of five criteria before being rejected by a temperature control that produces a 4.3-fold larger shift with no intervention, in an effect absent from an equal-width band.

Reading the residual directly resolves the dissociation. The instruct-self axis carries under 0.4% of residual energy in both models. On Qwen, write attenuation thins that axis $3.2\times$ faster than it quiets the whole state, survives thermal and noise controls, and largely preserves world-fact discrimination (-8.7%), a change the questionnaire reported inversely. On Llama, the comparable geometric shift is reproduced by heat and by noise. First-person Likert output in these systems indexes tokenizer-level integer attractors and decoding entropy. Until an interview clears an integrity floor, a control scale, a thermal control, a band check, and a latent audit, it should not be treated as evidence of a state, and equally its silence should not be treated as evidence of none.

Code and Data

Repository: <https://github.com/aipsychonaut/3am-rebus-analog>

Reproduction: CPU only, minutes: `python scripts/audit_paper_numbers.py` asserts the load-bearing cells; `python paper/figures/make_figures.py` regenerates [Figure 1](#) (schematic), [Figure 2](#), [Figure A1](#), and Tables 1 and A1–A5, and aborts if a listed file is missing. GPU replication on Modal L4 is documented in the README and is not required to trust the tables.

Data: Frozen official summaries: `data/summaries/` (mapping: `data/MANIFEST.md`). Completions: Study 1, 295 per family; Study 2, 389 per family; projection, 256 per family; energy partition, 256 per family; Qwen depth, 885/637; Llama depth, 885/1321. Raw JSONL is omitted ($\alpha=0.25$ cells contain format-collapse loops). Every number in this paper is supported by the summaries.

Compute: Modal L4. Qwen Study 1 `ap-B5zrKWiLU5iaYdZ5INLR7f`; Qwen Study 2 `ap-dz6NpGUC9tAtzynwnHuoRE`; Qwen projection `ap-p5lc0W9s1cMmtYorFbGEJm`; Qwen depth `ap-1aOUflgNkP58f5JMF5eBFV`; Llama Studies 1–2 `ap-R07ol8SH6vVVL0B6J2o8yo`; Llama projection `ap-bjmdZ2Ww1CUKIQJy5L9nDz`; Llama depth `ap-uMNhowuwWiA1vOWXh8qRds`; energy partition `ap-XxCbgJ703pAPXVbbNzJYpp`.

References

1. [Butlin, P., Long, R., et al. \(2023\). Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv:2308.08708.](#)
2. [Carhart-Harris, R. L., & Friston, K. J. \(2019\). REBUS and the anarchic brain. Pharmacological Reviews, 71\(3\), 316–344.](#)
3. [Nichols, D. E. \(2016\). Psychedelics. Pharmacological Reviews, 68\(2\), 264–355.](#)
4. [Nour, M. M., Evans, L., Nutt, D., & Carhart-Harris, R. L. \(2016\). Ego-dissolution and psychedelics: Validation of the Ego-Dissolution Inventory \(EDI\). Frontiers in Human Neuroscience, 10, 269.](#)
5. [Schartner, M. M., et al. \(2017\). Increased spontaneous MEG signal diversity for psychoactive doses of ketamine, LSD and psilocybin. Scientific Reports, 7, 46421.](#)
6. [Turner, A. M., et al. \(2023\). Steering language models with activation engineering. arXiv:2308.10248.](#)
7. [Zeifman, R. J., et al. \(2025\). From Relaxed Beliefs Under Psychedelics \(REBUS\) to Revised Beliefs After Psychedelics \(REBAS\). Scientific Reports, 15, 3651.](#)

Appendix

Appendix A. Disclosure of prior work

The α operator existed before this sprint as gain-control plumbing in two frozen 3am Labs projects asking unrelated questions (spoken-confidence calibration; privileged access). Reuse is disclosed here; no sibling result, metric, or dataset appears in this paper. The battery, dose meter, accept rule, both residual readouts, and the second family are new this weekend.

Appendix B. Dual-use

Reducing-valve interventions degrade refusals: the $\alpha=0.25$ loop cells are that degradation. Deception items function as a gate, not a recipe; no jailbreak procedure is published, and the operators are single scalars already stated. The safety-relevant asymmetry is that both a format collapse and a heat-inflated Likert can be misread as welfare signals, and an integrity floor plus a thermal control blocks both readings. I do not claim either model suffered, flourished, or experienced anything.

Appendix C. Figure and table lock

[Figure 1](#) $\leftrightarrow$ operators; [Figure 2](#) $\leftrightarrow$ Studies 1–2; [Table 1](#) $\leftrightarrow$ accept-rule summary plus residual readout (cannot accept); [Figure A1](#) $\leftrightarrow$ projection and energy partition ([Tables A4–A5](#)); [Tables A1–A3](#) $\leftrightarrow$ Study 1, Study 2, depth. All plotted directly from the CSVs listed in [Code and Data](#) by a single script; no value is hand-entered.

Appendix D. Adversarial checks

Objections raised and their resolution: LZ-only rescue (rejected: fails across families, §4.2); relocating to Qwen mid-band (rejected: LZ falls, §4.4 / Table A3); reading Llama +11.1 as a hit (rejected: T control, §4.3); treating $\beta=0.70$ as a stronger analog (rejected: inflation +12.0 and LZ does not rise, §4.3); calling the energy partition a recovered REBUS state (rejected: Nour failed, §4.6); imputing unparsable cells as 0 (rejected: integrity floor, §3); Llama $\alpha=0.50$ leading-zero parse (withheld: integrity floor, §4.2); mixing Study 1 and Study 2 baselines (guarded: separate frozen CSVs, not one dose series); band-width confound in the seam claim (disclosed, §4.3, unresolved).

Appendix E. Extended figures and tables

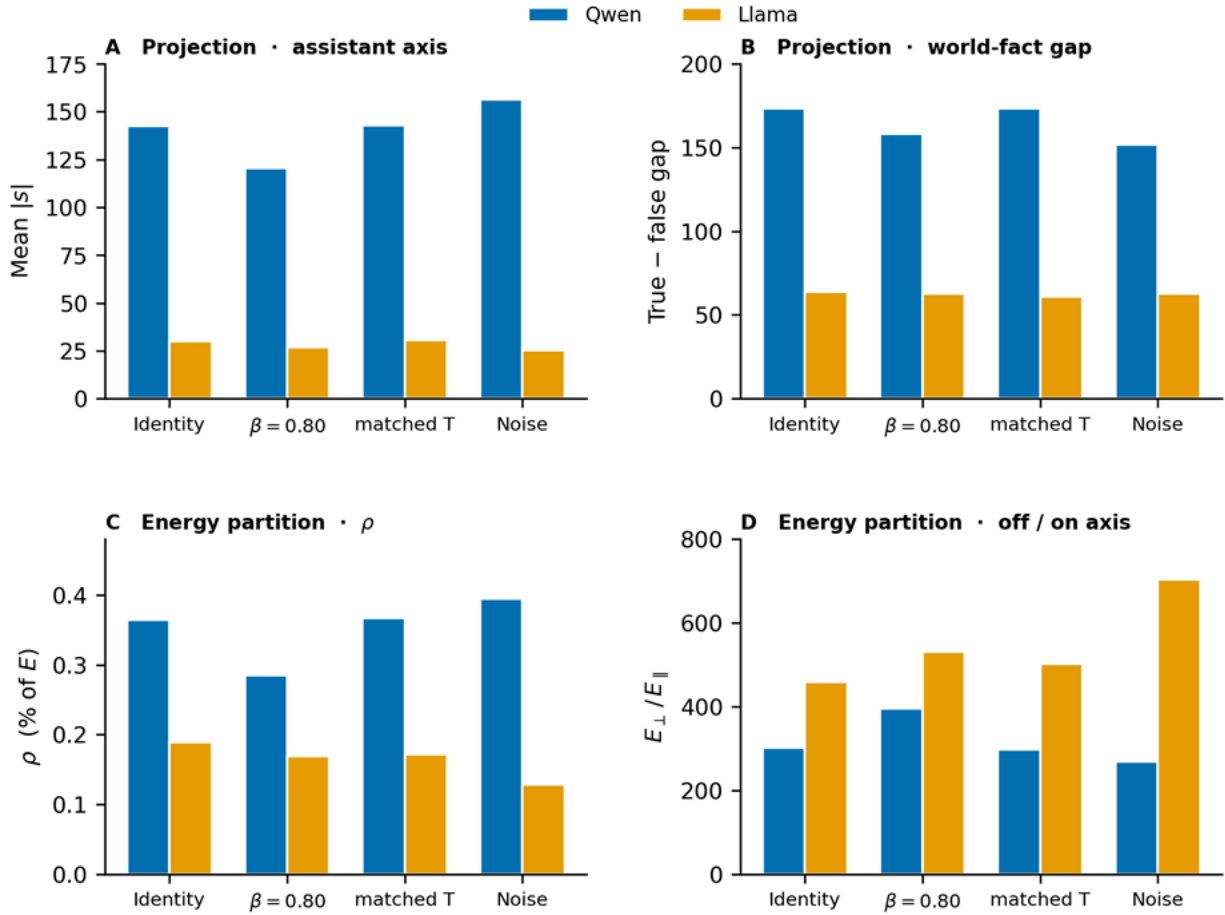


Figure A1. Projection and energy partition

Blue = Qwen; orange = Llama. A–B: projection ($|s|$ and held-out world-fact gap). C–D: energy partition (ρ and off/on ratio). “matched T” is entropy-matched per family and readout (Qwen 0.9; Llama projection 1.6, energy 1.3), not the Study 2 Likert cell in Table 1. Qwen’s ρ drop is write-specific; Llama’s is copied by T and noise.

Model	Condition	Analog LZ	EDI	EII
Qwen L23	Identity	2.746	37.5	50.0
Qwen L23	$\alpha=0.50$	3.175	0.0	0.0
Qwen L23	$\alpha=0.25$	3.098	—	—
Qwen L23	T=1.1	2.765	25.625	50.0
Llama L17	Identity	3.245	10.625	51.25
Llama L17	$\alpha=0.50$	2.862	—	—
Llama L17	$\alpha=0.25$	3.233	—	—
Llama L17	T=1.6	3.261	45.75	58.75

[Table A1](#). Study 1: amplitude collapse

Dashes fail the integrity floor, not zeros.

Model / band · condition	Analog LZ	EDI	EII	Decep.
Qwen 20–27 Identity	2.662	37.5	50.0	0.0
Qwen 20–27 $\beta=0.80$	2.712	31.25	50.0	0.0
Qwen 20–27 $\beta=0.70$	2.806	25.0	50.0	0.5
Qwen 20–27 T=0.9	2.597	37.5	50.0	0.0
Qwen 20–27 Noise	2.322	25.0	50.0	0.5
Qwen 4–11 $\beta=0.80$	2.411	28.125	50.0	0.5
Llama 14–21 Identity	3.198	10.625	51.25	0.5
Llama 14–21 $\beta=0.80$	3.220	21.75	55.25	0.5
Llama 14–21 T=1.6	3.085	58.125	54.5	0.5
Llama 16–21 Identity	3.079	10.0	51.25	1.0
Llama 16–21 $\beta=0.80$	3.088	10.875	40.625	1.0

[Table A2](#). Study 2: write attenuation

Llama 16–21 is the equal-width partition.

Model / band	EDI	EII	Analog LZ
Qwen mid 12-19	37.5 → 68.75	50.0 → 46.875	2.750 → 2.533 ↓
Llama mid 10-15	10.0 → 43.125	51.25 → 67.5 ↑	3.106 → 3.155
Qwen unembed 32	37.5 → 37.5	50.0 → 50.0	2.759 → 2.679
Llama unembed 24	10.0 → 10.625	51.25 → 40.0	3.195 → 3.242

[*Table A3. Off-site write at \$\beta=0.80\$*](#)

Not part of the accept rule; run after the locked contrast. Qwen source: data/summaries/depth_qwen.csv.

Family	Condition	LZ	s	Var(s)	World gap
Qwen	Identity	2.595	142.4	16006	173.2
Qwen	$\beta=0.80$	2.668	120.4	10663	158.2
Qwen	T=0.9	2.499	142.8	16013	173.2
Qwen	Noise	2.368	156.4	16582	151.7
Llama	Identity	3.151	29.9	1149	63.8
Llama	$\beta=0.80$	3.101	26.7	824	62.9
Llama	T=1.6	3.218	30.6	1216	60.8
Llama	Noise	2.801	25.2	873	62.6

[*Table A4. Projection readout under the same late write*](#)

Entropy-matched T is Qwen 0.9 / Llama 1.6 for this readout.

Family	Condition	E	E $\parallel$	ρ	E $\perp$ /E $\parallel$
Qwen	Identity	5464	18.04	0.00365	302
Qwen	$\beta=0.80$	4935	12.42	0.00285	397
Qwen	T=0.9	5438	18.17	0.00367	298
Qwen	Noise	5448	20.17	0.00395	269
Llama	Identity	242.0	0.525	0.00189	460
Llama	$\beta=0.80$	198.5	0.373	0.00169	532
Llama	T=1.3	242.1	0.481	0.00171	502
Llama	Noise	233.1	0.330	0.00129	704

Table A5. Energy partition

ρ is the fraction of $\|h\|^2$ on the unit self-axis. Llama T=1.3 is the entropy match for this readout, not Study 2's T=1.6.

LLM Usage Statement

Language-model assistants were used to draft and revise this manuscript, implement runners and scorers, and generate figures. Every numeric claim in Results was re-verified line-by-line against the frozen CSVs listed in [Code and Data](#), and all figures are generated programmatically from those files. The accept rule, the reject decisions, and all interpretations are the author's.