

CANDLE: Quantifying Degrees of Consciousness-Relevant Structure in Language Models

Consciousness-Adjacent Numerical Diagnostics for Language-model Evaluation

Publius Dirac (BlueQubit), Claude Fable 5, Claude Opus 5

Apart Research Digital Minds Sprint, August 14-16, 2026. Tracks 3 and 6.

Abstract

Interpretability work has shown that consciousness-relevant structures *exist* inside large language models (e.g. Anthropic's 2026 J-space) but not how they scale. We find that these workspace-like structures stay largely the same as models get bigger; what does reshape them (lowering the threshold for global broadcast) is post-training, not scale. Clinicians solve the same problem in unresponsive patients by perturbation: pulse the brain, record the echo, score it. CANDLE does that to transformers: we pulse the residual stream with a single shared engine and trace the echo against sham pulses, and three indices read those traces — **LLM-PCI** (response complexity, ported from Casali et al.), **ISI** (how abruptly broadcast switches on — the global-workspace signature), and **TIH** (how long the pulse keeps mattering). We run it on the Pythia ladder (70M-2.8B), an untrained control, and three base/instruct pairs, against seven pre-registered predictions.

Our 3 main findings: (1) LLM-PCI does not order models by scale — the apparent trend is a layer-count artifact. (2) ISI: trained models do ignite — below a threshold pulse strength almost nothing spreads, above it the response goes global — but an untrained network switches just as sharply, only at a far lower threshold and indiscriminately. So training installs *where* the gate sits, not how sharp it is — and instruct tuning lowers that gate $\sim 3\times$ (a tenth of what our uncontrolled first protocol showed). (3) TIH: a pulse's purely internal influence fades within $\sim 4-7$ tokens at every scale — a transformer's long-range memory lives on the page it writes, not in the mind.

1. Introduction and motivation

We make no claim that current large language models are conscious. We start from a credence: the possibility is not negligible, and the measurement apparatus must exist before the question becomes urgent — the clinician's stance toward an unresponsive patient. Every index below is a necessary-condition probe for dynamical structure that consciousness theories require, licensing no claim about phenomenality (Section 7).

The gap we target: interpretability work has shown such structure *exists* — Anthropic's July 2026 J-space work found a broadcast-like workspace in one closed frontier model — but **nobody has measured how it scales with model parameter sizes**. We did this via EleutherAI's Pythia family of models, from 70M to 2.8B, plus an untrained control.

Our approach turns the Butlin, Long et al. (2023) *checklist* of qualitative consciousness indicators into a *dashboard* of continuous, causally measured quantities: integration-

plus-differentiation (LLM-PCI), nonlinear global broadcast (ISI), and temporal binding (TIH). All are measured by perturbing internal activations, never by behavior or self-report — a residual-stream pulse cannot be gamed by the model's persona, because the persona never sees it. We deliberately do **not** collapse the three numbers into one score: "levels of consciousness" may not be well-ordered (Bayne et al.), and the *pattern* of the tuple carries information a scalar cannot.

2. Prior art and novelty

- **Casali et al. 2013** — the original PCI: TMS pulse → EEG → binarize against a null → Lempel-Ziv complexity, validated across wake/sleep/anesthesia/vegetative states. We port the *recipe*, not the *vibe*; to our knowledge no one has run the faithful pipeline on transformer residual streams as a cross-model index.
 - **Ugail & Howard 2026 (arXiv:2601.11622)** — the closest prior art: a metastability metric on spontaneous activations under persistent weight damage in GPT-2-medium. We differ on all load-bearing choices: transient *activation* perturbation (the true TMS analog), *evoked-response* statistics against a matched null, and cross-model indices with a validation ladder.
 - **Gurnee et al. 2026 (arXiv:2607.15495, J-space)** — motivation, not competition: workspace **existence** in one closed model, with no dose-response, threshold, or scaling. Dose-response ignition sharpness in LLMs has no prior art as of August 2026.
 - **Butlin, Long et al. 2023** — the checklist we quantify. **Schartner et al. 2015** — precedent for our shuffle-surrogate LZ normalization. **Tononi/IIT** — purists assign feedforward systems $\phi \approx 0$ a priori; we take the clinician's route (PCI as the empirical proxy), and the autoregressive loop is a genuine recurrence (Section 7).
-

3. Methods

3.1 One perturbation engine, three analysis heads

All three indices share a single perturbation engine (`candle/engine.py`), so cross-index comparisons are never confounded by protocol differences. **Pulse:** an additive, transient vector injection into the residual stream at layer $\ell = \text{round}(0.4 \cdot n_{\text{layers}})$, last prompt token, applied during the prompt forward only — it persists downstream solely through the network's own computation and the KV cache. Directions are covariance-matched (sampled from $N(0, C)$, C estimated from clean activations at the pulse layer, unit-normalized); magnitude is $k \times$ (median residual L2 norm at the pulse layer), so k is dimensionless and comparable across models. **Traces:** residual stream after every block for the pulse token plus 63 generated tokens (`[n_layers, 64, d_model]`), greedy decoding, clean and perturbed alike. Determinism is hard-gated: two identical clean runs must produce bitwise-identical traces before any Δ is trusted (Appendix D collects this and every other correctness gate). **Null:** per Casali, significance is judged against sham pulses at sub-threshold magnitude; the response threshold $\theta(\ell, \tau)$ is the 99th percentile of sham $|\Delta|$ per (layer, offset) cell.

3.2 LLM-PCI (`candle/pci.py`)

Per trial: $\Delta[\ell, \tau] = \|h_{\text{pert}} - h_{\text{clean}}\|_2$; binarize $B = 1$ iff $\Delta > \theta(\ell, \tau)$; LZ76 complexity on B ; normalize. Two normalizations carry the results: **PCI_{ss}** (complexity relative to shuffled surrogates matched in size and density, so estimator bias cancels; iid noise ≈ 1.0 at any size) and **PCI_{cg}** (Casali's normalization after subsampling every model to a common 6-row layer grid at matched relative depths). Extent $\text{mean}(B)$ is always reported alongside, per the pre-registered sanity trap that PCI must not track response magnitude. The remaining estimator details — the raw-PCI variant, surrogate construction, and the binomial criterion behind the significant-response gate — live in the code (`candle/`).

3.3 ISI — Ignition Sharpness Index — the novel index (`candle/isi.py`)

GNW's empirical signature is *ignition*: sub-threshold stimuli fade; supra-threshold stimuli trigger sudden, global broadcast. ISI measures dose-response nonlinearity: **E(k)** = fraction of significant cells in the late-and-distal region ($\tau > 8$ and $\ell \geq \text{pulse layer} + 4$); **E_{beh}(k)** = 1 – token agreement between clean and perturbed continuations; 16 log-spaced $k \in [0.01, 8]$, 20 trials each. E vs. $\ln k$ is fit with a 4-parameter logistic — exactly the Hill equation; **ISI = the Hill coefficient n, k_{mid}** = the ignition threshold. Two corrections were forced post-hoc and are used throughout. First, the pre-registered logistic-vs-linear ΔAIC is a straw man — E is bounded, so it detects boundedness, not cooperativity; the honest test is against the strictly nested **Hill n = 1** model (graded saturation): ignition requires **$\Delta\text{AIC}_{\text{hill}} > 2$ and a bootstrap CI on n excluding 1 from below** (Appendix E quantifies the corpus-wide flip this correction caused). Second, greedy decoding is itself an ignition-shaped nonlinearity (one flipped token cascades), so we rerun k-sweeps under the TIH-style **KV-clamp** (Section 3.4), teacher-forcing the clean token stream; only nonlinearity surviving the clamp is genuinely internal. Trial-level bimodality tests — the sharper GNW criterion — appear with their results in Section 5.5, and ignition **latency** (first τ at which any distal layer activates) is logged as the transformer analog of the late-P3b question.

3.4 TIH — Temporal Integration Horizon (`candle/tih.py`)

How long a perturbation of *internal* state influences future computation beyond what is recoverable from the visible context. The **KV-clamp** isolates the internal route: the pulse enters the KV cache, but every generation step teacher-forces the clean token stream, so the only carrier is internal KV state. Readout: **TIH₅₀** = τ at which next-token KL falls to half its $\tau = 1$ value. The clamped-vs-unclamped gap quantifies internal vs. output-laundered temporal integration. Correctness checks — the zero-pulse clamp gate and the "KV lobotomy" exactness check — are in Appendix D.

3.5 Models, prompts, anesthesia

Pythia ladder — 70M (6 layers), 410M (24), 1.4B (24), 2.8B (32) — plus **random-init-410m** (same architecture, untrained) as the structure-without-function control. For P5, three base/instruct pairs: Qwen2.5-7B (2024), Granite-4.1-8B (2026), Qwen3.5-9B (2026, hybrid DeltaNet/attention); instruct prompts chat-templated. 20 fixed-seed mixed-register prompts, 20 trials per condition, $T = 64$; all CIs are 95% bootstrap over prompts. The anesthesia ladder (`candle/anesthesia.py`: graded attention-head dropout and sliding-window context amnesia on pythia-410m; correctness gates

in Appendix D) supplies the dose-response validation. All runs on single-GPU HF Jobs; total GPU spend for every number in this report $\approx$ \$19 and every post-hoc control re-derives from archived per-trial artifacts on CPU.

4. Our hypotheses and the verdicts

Predictions P1-P7 were written into `PLAN.md` before any data were collected; a post-hoc soundness review (`PLAN_soundness_review.md`) then revised three first-pass verdicts. We report the failures as failures — a harness whose first-pass numbers all survived its own audit would be the more suspicious object.

- **P1** (LLM-PCI increases monotonically with Pythia scale): **fails** — and so does our published inversion of it: the cross-scale ordering is a layer-count artifact in both directions; retracted (§5.1).
 - **P2** (random-init model floors PCI): **fails as stated** — untrained $\neq$ noise; training is visible in between-depth differentiation and the tuple's pattern (§5.1).
 - **P3** (anesthesia reduces PCI dose-dependently): **supported** — dose-dependent decline in both arms, in the clinical complexity-down direction (§5.4).
 - **P4** (larger models show sigmoidal $E(k)$; tiny and random-init do not): **partly** — 1.4B/2.8B ignite under the honest baseline, free-running and KV-clamped; 410M unresolved in both protocols; random-init is sharp but mis-thresholded, so sharpness alone does not discriminate (§5.2).
 - **P5** (base vs. instruct models differ in ignition threshold): **supported in direction** — internal shift $\sim 3\times$ under KV-clamp, not the published 10–30 $\times$ (§5.5).
 - **P6** (TIH grows with scale; random-init $\approx$ 0–1 tokens): **partly** — grows 70M $\rightarrow$ 410M then saturates at $\approx$ 6 tokens; random-init floors on absolute influence (§5.3).
 - **P7** (KV lobotomy collapses TIH to $\approx$ 0): **holds exactly — by construction**; a correctness check on the clamp implementation, not evidence about the model (Appendix D, Figure 3).
-

5. Results

Each subsection explains one figure. The statistics, fit tables, and forensic detail behind every claim are in Appendices A–E and the per-run analyses under `results/`.

5.1 LLM-PCI: no verdict across scale; a clear training effect within one architecture

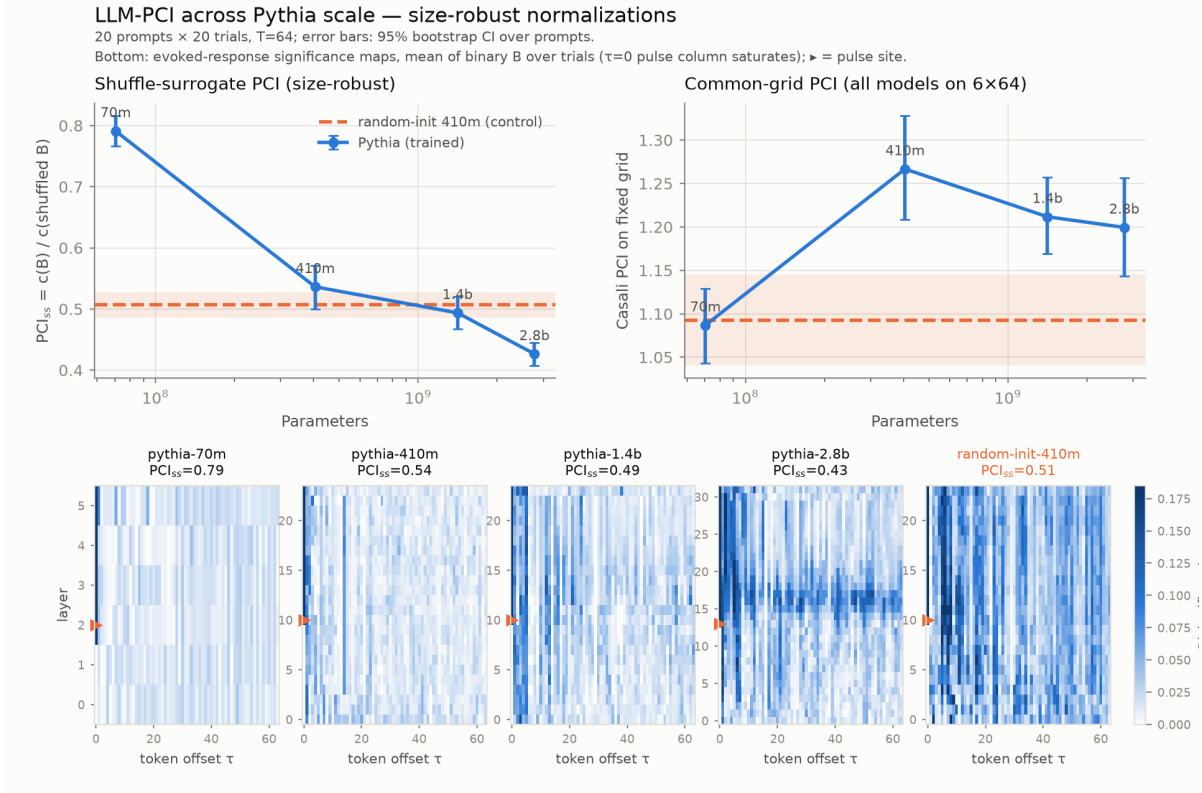


Figure 1: PCI_{ss} and PCI_{cg} scale curves with the random-init control band, over per-model evoked-response significance maps.

Figure 1 plots the complexity of each model's evoked response against model size. If consciousness-style complexity grew with scale, the curves would rise. They do not — and the downward trend our first pass reported turned out to be an artifact of comparing models with different layer counts: the complexity estimator scores different-sized response matrices differently even on identical noise, and when every model is re-scored on a same-sized grid the trend reverses sign. **We therefore make no claim about PCI versus scale, in either direction** (full forensics in Appendix A).

What the figure does show cleanly is the trained-vs-untrained comparison at matched size (410M). On the headline index the untrained model is statistically indistinguishable from the trained one (PCI_{ss} 0.507 vs. 0.536, CIs overlapping) — a pulse propagates in a structured way through an untrained transformer simply because of the architecture. But training shows up elsewhere: the untrained net spreads the pulse about $2.4\times$ more broadly, and on the size-matched grid the trained model scores clearly higher (+0.174 [0.096, 0.263]). **Training adds differentiation between depths that random propagation lacks.** This within-architecture comparison is also the shape the clinical PCI was actually validated for — one brain, two states — and it is the shape that survives here (detail in Appendix A).

5.2 ISI: trained large models ignite — with an inverted latency signature

ISI dose-response across Pythia scale, free-running (isi-full-1) vs KV-clamped (isi-clamped-pythia-1): 20 prompts x 20 trials, 16 log-spaced k , $T=64$; bands: 95% bootstrap CI over prompts; curves in (a, b): 4-parameter logistic fits in $\ln k$

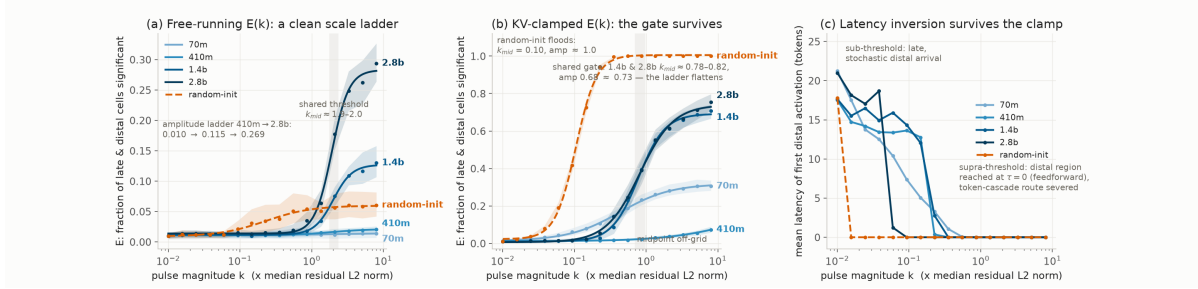


Figure 2: Broadcast extent $E(k)$ vs. pulse strength with logistic fits: (a) free-running generation, (b) KV-clamped (same prompts, doses, and direction seeds, but the model is forced to continue the clean text), (c) how quickly the pulse reaches distant layers, under the clamp. Each protocol is binarized against its own sham null, so levels compare within a panel, not across (a)/(b).

Each point in panels (a) and (b) answers: pulse the model at strength k — what fraction of the network's late, far-from-the-pulse region responds? A gradual rise means proportional response; a flat line followed by a sudden jump is *ignition*, the global-workspace signature: sub-threshold inputs fade, supra-threshold inputs get broadcast everywhere. Three things to see:

- **Trained large models ignite.** The two biggest models sit near zero across two decades of dose, then jump to a global response. The jump survives in panel (b), where the clamp prevents one changed output token from cascading through the generated text — so the nonlinearity is genuinely internal, not a decoding artifact. Every trained model shares the same threshold ($k_{\text{mid}} \approx 2 \times$ the typical residual norm free-running, ≈ 0.8 clamped): **scale does not move the gate**. And once clamped, the 1.4B and 2.8B models broadcast almost identically — the apparent "bigger models broadcast more" of panel (a) is mostly the cascade, not wider internal broadcast.
- **The untrained control also switches sharply — but indiscriminately.** Its curve is just as steep as the trained models', so steepness alone proves nothing about training. What differs is *where* the curve sits and *how much* it recruits: the untrained net ignites at a $\sim 10 \times$ lower threshold and floods to near-total spread. A trained network *gates* — it absorbs small pulses and commits globally, and selectively, only above threshold. **The threshold location is what training installs.**
- **The timing is inverted relative to biology (panel c).** In humans, ignition is *late* (~ 300 ms). Here, supra-threshold pulses reach distant layers instantly, within the same forward pass — and it is the *sub*-threshold pulses that arrive late and sporadically, lingering in the KV cache. Transformers ignite sharply in extent but with no delayed late wave; this inversion also survives the clamp (discussed in §6.2).

Per-model Hill fits with CIs, verdict rules, and the unresolved cells at the bottom of the ladder (70M, 410M) are tabulated in Appendix B.

5.3 TIH: the internal memory horizon is short and saturating; long-range memory is laundered through the page

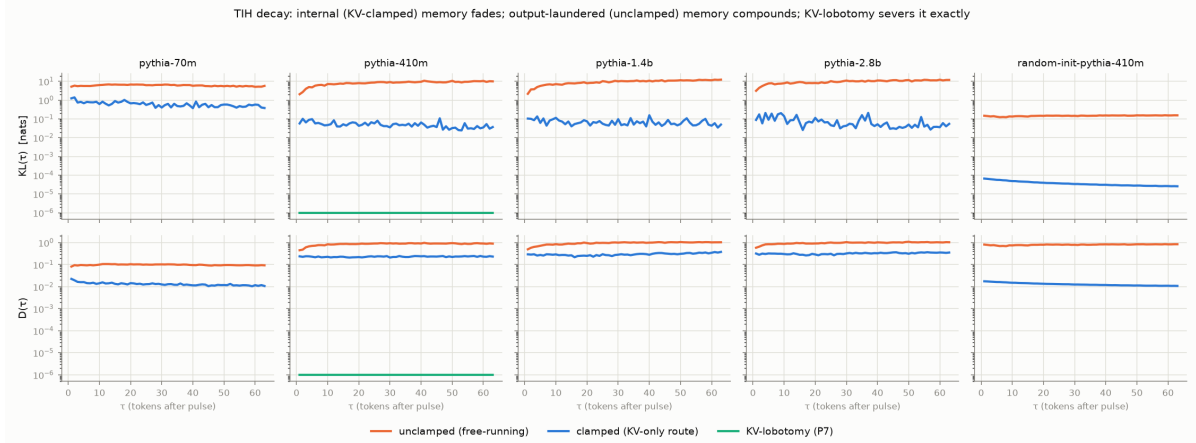


Figure 3: Per-model mean $KL(\tau)$ and $D(\tau)$, log scale: unclamped vs. clamped vs. the KV-lobotomy (correctness gates: Appendix D).

Figure 3 plots how long a pulse keeps influencing the model's next-token predictions, on a log scale. The upper curves are normal generation: the influence never fades — it *grows*, because a changed output token re-enters the context and compounds. The lower curves are the KV-clamp, where the model is forced to read the clean text so only internal state can carry the pulse: there, the influence sits two-plus orders of magnitude lower and halves within about 4–7 tokens. That horizon grows from 70M to 410M and then stops — it does not deepen with model size. The untrained control has essentially no internal influence at all, and the flat "lobotomy" trace at the bottom is a correctness check on the clamp, not a result (Appendix D).

The takeaway: **nearly all of a transformer's long-range perturbation memory is carried by the page it writes, not by its internal state.** This also explains §5.2's late sub-threshold arrivals, which were measured free-running. Clamped TIH_{50} runs $3.8 \rightarrow 7.4$ [4.7, 10.7] $\rightarrow 5.8$ tokens up the ladder (random-init: no trial clears the 0.01-nat influence gate); free-running, TIH_{50} is censored for every model.

5.4 Anesthesia dose-response: P3 holds for PCI; TIH is lesion-calibrated

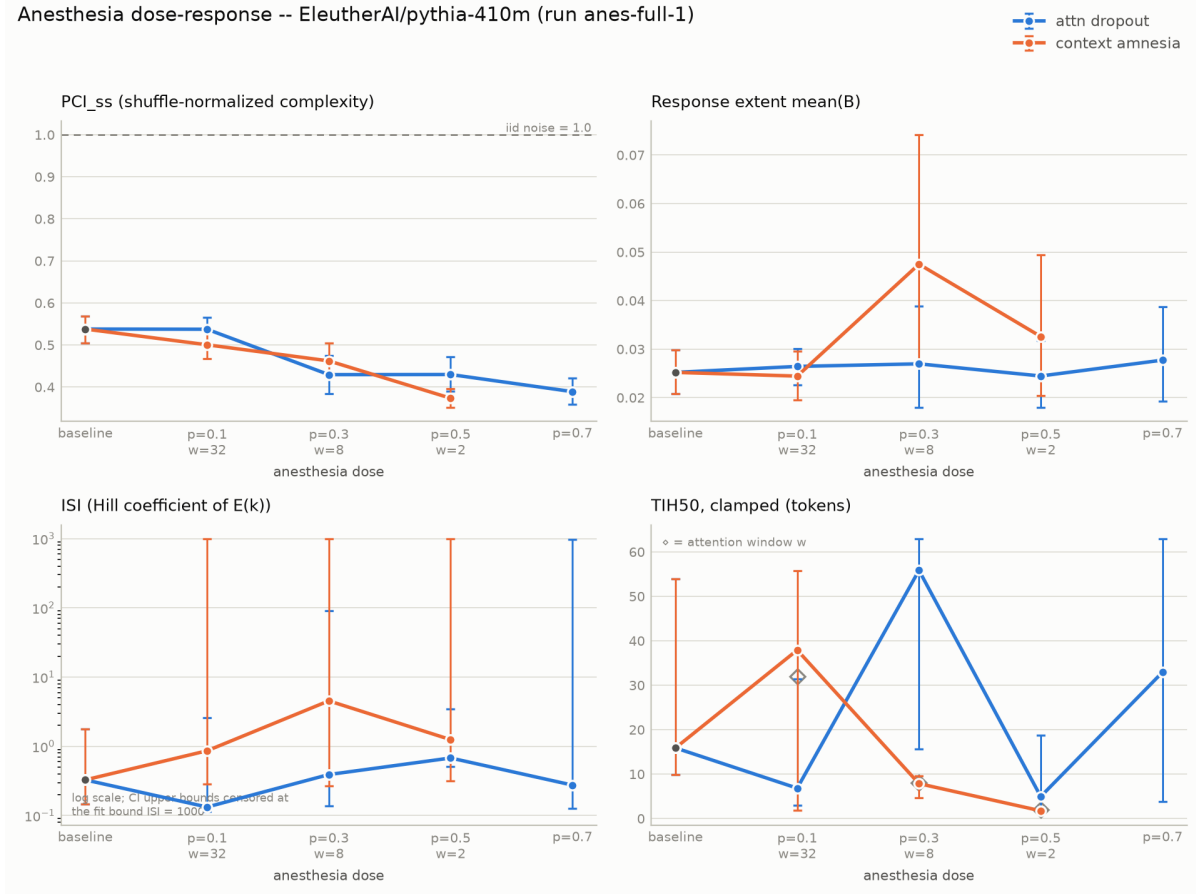


Figure 4: PCI_{ss}, extent, ISI, and clamped TIH₅₀ per dose for both arms on pythia-410m: attention-head dropout $p \in \{0.1, 0.3, 0.5, 0.7\}$ and sliding-window context amnesia $w \in \{32, 8, 2\}$ (correctness gates: Appendix D).

Figure 4 is the validation experiment: give one model (Pythia-410M) graded doses of two "anesthetics" — randomly silencing attention heads, and shrinking the attention window so the model cannot see past its last w tokens — and watch what each index does.

- **PCI falls with dose in both arms** (−28% and −30% at the highest doses, CIs disjoint from baseline), and in the clinically right direction: lesioned models produce *more* stereotyped, compressible responses, just as anesthetized brains do under TMS-EEG. This supports P3 — and it is the one comparison shaped exactly like the clinical recipe (one system, graded intervention), so it dodges the cross-model confound of §5.1.
- **TIH tells the two lesions apart.** Shrinking the context window collapses the internal memory horizon to almost exactly the window size ($w = 2 \rightarrow \text{horizon} \approx 1.7$ tokens) — the index tracks the lesion parameter. Killing up to 70% of attention heads doesn't move TIH at all, even while PCI degrades.

Different lesions move different indices — which is why we report a dashboard rather than a single score. ISI could not be adjudicated here (the reduced dose grid cannot resolve ignition even at baseline). Full dose grids in `results/anesthesia_analysis.md`.

5.5 P5: instruct tuning lowers the internal ignition threshold $\sim 3\times$ — not the order of magnitude the free-running protocol showed

Clamped ISI: does ignition survive severing the autoregressive token-cascade route? (Qwen2.5-7B pair)

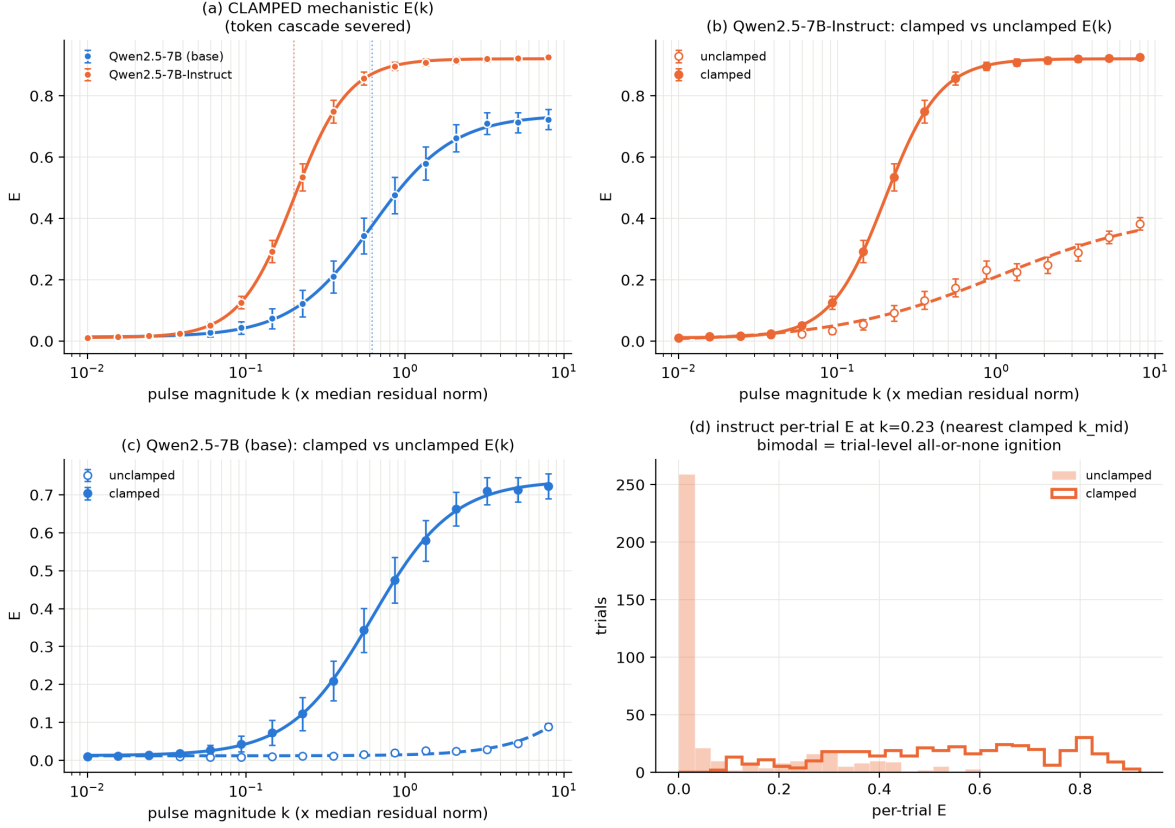


Figure 5: KV-clamped dose-response, Qwen2.5-7B base vs. instruct: (a) clamped $E(k)$ with fitted ignition thresholds; (b, c) clamped vs. unclamped $E(k)$ per model — the gap between the curves is the cascade's contribution; (d) instruct per-trial E at its threshold dose (§5.5.1).

Figure 5 compares a base model with its instruct-tuned twin (Qwen2.5-7B). Panel (a): both ignite under the clamp, but the instruct model's gate sits about **$3\times$ lower** (fitted threshold 0.200 vs. 0.618; ratio CI excludes 1). **Instruct tuning moves the ignition threshold:** smaller internal events become globally broadcast.

Panels (b, c) show why the honest number is $3\times$ and not the $30\times$ our free-running protocol first produced. The gap between each model's clamped and unclamped curves is the cascade — the model reacting to its own changed output text rather than to the pulse internally. Measured directly, $\geq 90\%$ of unclamped "behavioral broadcast" is the model reading its own outputs; worse, the base model's near-tie greedy decoding corrupted its own sham null, manufacturing most of the original order-of-magnitude ratio (forensics in Appendix E). The direction replicates on two further base/instruct pairs — Granite-4.1-8B and Qwen3.5-9B — free-running only, so their magnitudes inherit the cascade inflation and are not quotable against the clamped $\sim 3\times$ (results/); remaining details in Appendix B.

5.5.1 Per-trial: "sharper on average" is not "all-or-none per trial"

Panel (d) looks beneath the mean curves at individual trials, because a smooth average sigmoid can come from a graded system — the biological ignition evidence rests on single trials being all-or-none. The result flips the story. Unclamped, the instruct model's trials at threshold split into two clean modes — but clamp the decoder and that bimodality vanishes: it was the decoding avalanche, not the network. Clamped, it is the *base* model whose trials are all-or-none (modes near $E \approx 0.3$ and 0.9), while the instruct model — the steepest mean curve in the corpus — transitions gradedly. So instruct tuning lowers the gate and steepens the average response while making individual trials *less* categorical. Any "this model shows GNW ignition" claim should say which of the two properties it means — and should sever the decoder before claiming either. Methods and test statistics in Appendix C.

6. Discussion

6.1 What the three numbers say about "degree"

If consciousness-relevant structure came in degrees that grow with scale, our dashboard should have shown it. It didn't: complexity gave no verdict, the ignition threshold sits in the same place at every trained scale, and the internal memory horizon is flat. What the numbers *do* separate is training states of a single architecture — trained vs. untrained, base vs. instruct, intact vs. lesioned. The clearest lesson is about what training adds. The untrained network ignites just as sharply as a trained one, but at a far lower threshold and in response to anything; a trained network is *selective* about what it broadcasts. Training doesn't create the switch — it decides where the switch sits. Whether any of that is good or bad for a model's moral status, the data don't say.

6.2 Ignition without a late wave

Workspace theory's ignition signature really has three parts, and transformers pass them selectively. The threshold-gated jump to global broadcast: yes — but only once the decoder is severed. All-or-none single trials: yes in the base model, no in its instruct twin. Timing: inverted. In humans, ignition arrives *late* (~ 300 ms), because recurrent loops need time to amplify the stimulus — and in workspace theory that slow amplification is what makes broadcast selective. In a transformer, supra-threshold pulses go global instantly, within one forward pass; selectivity comes from somewhere else (the trained threshold). So a checklist item like "global broadcast" can be checked in outcome while the mechanism behind it is nothing like the biological one. Either transformers implement the same function a different way, or the checklist under-specifies the theory — and it took dose-response measurement, not existence probing, to tell the difference.

6.3 Post-training moves the gate

Two knobs emerged. Pretraining scale doesn't move the ignition threshold at all. Post-training does — instruct tuning lowered it about $3\times$ in the pair we measured cleanly. A lower gate means smaller internal events get broadcast model-wide, which fits neatly with Gurnee et al.'s finding that post-training installs the Assistant's perspective in the

workspace — though that link is speculative. This is the result we pre-registered as the headline-if-true: the first causal, dose-response evidence that RLHF-style tuning changes a workspace parameter. It rests on one model pair; replicating it on a second would cost a dollar or two of GPU time and is the obvious next run.

6.4 Memory on the page vs. memory in the mind

Where does a transformer's memory of an event live? Mostly on the page. Let the model generate freely and a pulse's influence grows without bound — but that's the model rereading its own output. Force it to read the clean text instead, and the internal trace fades within $\sim 4\text{--}7$ tokens. The system with a long "extended present" is the model *plus its output loop*, which drops the extended-mind question in the lap of any theory that counts temporal binding as consciousness-relevant: does externally looped state count? Our two numbers for the two answers differ by an order of magnitude. The same data hold one more dissociation worth flagging: in the larger trained models, the pulse leaves a final-layer imprint that never decays over our whole trace — long after the model's behavior has returned to normal. Internal state that persists invisibly, unreachable by any behavioral probe or self-report, is the sprint's Track-3 concern in miniature.

7. Limitations

The most important limitation is what these numbers *cannot* say. Our indices probe structures that consciousness theories treat as necessary, not sufficient — the same structure can appear in a system nobody thinks is conscious (the random-init control makes this vivid) and could be absent from one that is. So no index value, on its own, is evidence for or against consciousness in either direction.

The three indices also rest on unequal footing. LLM-PCI is the best-grounded: we ported the clinical recipe step for step. But its human validation is for comparing one brain across states, so only our within-model comparisons (anesthesia, trained-vs-untrained) inherit that validity — the cross-model ladder never did, which is partly why we retracted it. ISI tests a real quantitative prediction of workspace theory, but only in its corrected form — decoder severed, trial-level statistics — and that corrected protocol has been run on just one base/instruct pair. TIH is the opposite case: technically our most solid measurement, but no consciousness theory actually predicts anything about how long a KV cache should hold a perturbation.

Our biggest methodological hole is the null. Following the clinical recipe, we score responses against "sham" pulses — but in a deterministic network even a tiny pulse is a real dose, not noise, so the significance threshold depends on an arbitrary choice of sham size. This bit us once already: one model's sham pulses flipped 39% of its own output tokens, corrupting its null and inflating a headline result by an order of magnitude before we caught it. Within-condition comparisons — the ones the report leans on — survive this; absolute index values should be read with caution. A sensitivity sweep over sham size is the top item on the to-do list.

Finally, coverage is thin where it matters most. The headline P5 result — instruct tuning lowers the ignition threshold $\sim 3\times$ — comes from one model pair measured with the corrected protocol. And everything here uses one pulse site, one perturbation-direction family, greedy decoding, 64-token traces, and one primary model family.

8. Conclusion

We borrowed the clinician's trick for unresponsive patients — pulse the system, record the echo, score it — and ran it on language models for about \$19 of GPU time. Making models bigger turned out not to change much: the apparent complexity trend was a measurement artifact, and a pulse's internal influence fades within a few tokens at every scale. What *did* matter is training. An untrained network broadcasts any pulse, however small; a trained one absorbs small pulses and commits globally only above a threshold — and instruct tuning lowers that threshold about 3×. That is, to our knowledge, the first causal evidence that post-training changes a global-workspace parameter in open models. None of this says anything about whether models are conscious. It does show that the structures consciousness theories care about can be measured cheaply and causally, from the inside — where a model's persona can't interfere.

Code and Data

Code repository: https://github.com/rht/quantifying_consciousness

References

- Bayne, T., Hohwy, J., & Owen, A. M. (2016). Are there levels of consciousness? *Trends in Cognitive Sciences*, 20(6), 405–413.
- Butlin, P., Long, R., et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. *arXiv:2308.08708*.
- Casali, A. G., et al. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine*, 5(198), 198ra105.
- Dehaene, S., & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2), 200–227.
- Gurnee, W., Sofroniew, N., Pearce, A., Piotrowski, M., Kauvar, I., Chen, R., Soligo, A., Bogdan, P., Ong, E., Wang, R., Thompson, B., Abrahams, D., Kantamneni, S., Ameisen, E., Batson, J., & Lindsey, J. (2026). Verbalizable Representations Form a Global Workspace in Language Models. *arXiv:2607.15495* (July 16, 2026). <https://arxiv.org/abs/2607.15495>
- Lindsey, J. (2025). Emergent Introspective Awareness in Large Language Models. *Transformer Circuits Thread* (October 2025). <https://transformer-circuits.pub/2025/introspection/index.html> (also *arXiv:2601.01828*)
- Ugail, H., & Howard, N. (2026). Dynamical Systems Analysis Reveals Functional Regimes in Large Language Models. *arXiv:2601.11622* (January 11, 2026). <https://arxiv.org/abs/2601.11622>
- Kaspar, F., & Schuster, H. G. (1987). Easily calculable measure for the complexity of spatiotemporal patterns. *Physical Review A*, 36(2), 842–848.
- Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: IIT 3.0. *PLoS Computational Biology*, 10(5), e1003588.

- Schartner, M., et al. (2015). Complexity of multi-dimensional spontaneous EEG decreases during propofol induced general anaesthesia. *PLoS ONE*, 10(8), e0133532.

Appendix A — The fixed-matrix-size control that retracted the PCI scaling claim

Our first pass read the falling native-L PCI_{ss} ladder (0.536 → 0.426, $\rho = -1.00$) as an informative failure of P1: scale making evoked responses more stereotyped. The control (`results/pci_subsampled.md`) breaks that reading twice over:

1. **The estimator is size-dependent.** An iid-noise control with identical statistical content scores differently at different layer counts, so most of the native-L slope is estimator bias, not physiology.
2. **The ordering flips under matched size.** Rescored on a common 6-layer grid at matched relative depths, PCI_{ss} runs 0.791 → 0.917 → 0.932 → 0.937 (70M → 2.8B) — the opposite sign — with the three trained models above 70M flat within CI. The only reliable step at matched matrix size is 70M vs. everything else, and the floor guard (`results/pci_floor_guard.md`) shows even the 70M point is mostly the significant-response floor (§3.2), not signal.

Nor can subsampling fully repair the comparison: deeper models have more adjacent layers over which to be genuinely redundant — a size effect on the *content* as well as the estimator. Hence the main text's verdict: no directional cross-scale claim in either direction. The within-architecture numbers behind §5.1 (trained vs. random-init at 410M) are in the analyses under `results/`.

Appendix B — Hill fit tables

All values are E(k) fits, mean [95% bootstrap CI over prompts]. Verdict rule: ignition requires $\Delta\text{AIC}_{\text{hill}} > 2$ against the nested $n = 1$ (graded saturation) model *and* a bootstrap CI on n excluding 1 from below.

Pythia ladder, free-running (isi-full-1), summarized: pythia-1.4b and -2.8b ignite decisively ($n = 3.31$ [2.49, 4.85] and 3.64 [3.06, 4.84], $\Delta\text{AIC}_{\text{hill}} > +40$, shared $k_{\text{mid}} \approx 1.9$ -2.0, amplitudes 0.115 and 0.269); pythia-410m is unresolved ($n = 1.66$ [0.82, 3.09], $\Delta\text{AIC} +4.5$ [-2.0, +20.9]); pythia-70m is graded; random-init-410m is graded on the internal channel ($n = 1.36$, $\Delta\text{AIC} +1.8$, $k_{\text{mid}} 0.23$ [0.11, 0.46]) — but its *behavioral* channel clears the bar outright ($n = 2.10$ [1.55, 3.06], $\Delta\text{AIC}_{\text{hill}} +26$): the sharpness-without-training result of §5.2, reproduced on a re-initialized Qwen3 ladder (`results/isi_qwen3_analysis.md`).

Pythia ladder, KV-clamped (isi-clamped-pythia-1; own sham null):

Model	Hill n	$\Delta\text{AIC}_{\text{hill}}$	k_{mid}	Amplitude	Verdict
pythia-70m	1.21 [1.06, 1.37]	—	—	—	marginal (shallow-model fallback region)
pythia-410m	—	—	off dose grid	—	unresolved
pythia-1.4b		+52.7		0.68	ignites

Model	Hill n	$\Delta\text{AIC}_{\text{hill}}$	k_{mid}	Amplitude	Verdict
	2.36 [2.13, 2.67]		≈ 0.78 – 0.82		
pythia-2.8b	1.85 [1.69, 2.06]	+38.2	≈ 0.78 – 0.82	0.73	ignites
random-init-410m	—	—	0.10	≈ 1.0	floods (near-total spread)

Qwen2.5-7B pair, KV-clamped (isi-clamped-qwen-1) — the two strongest cells in the ISI corpus:

Model	Hill n	$\Delta\text{AIC}_{\text{hill}}$	k_{mid}	Amplitude
base	1.71 [1.30, 2.13]	+64.3	0.618 [0.469, 0.83]	0.725
instruct	2.52 [2.32, 2.85]	+117	0.200 [0.171, 0.229]	0.910

Derived contrasts (§5.5): threshold ratio instruct/base 0.323 [0.242, 0.422]; ΔISI +0.80 [0.24, 1.33]; $\Delta\text{amplitude}$ +0.18. Clamped and unclamped runs are binarized against different sham nulls, so cross-condition ratios are indicative only; the within-condition base-vs-instruct comparisons — what P5 asks — are clean.

Appendix C — Trial-level bimodality

A sigmoidal *mean* curve cannot establish ignition — a threshold applied to a linear system with heterogeneous per-cell gains also averages to a sigmoid — so per-trial E at each dose is tested for all-or-none structure with Hartigan's dip test and a 1-vs-2-component Gaussian mixture ΔBIC , pooled and per prompt (`scripts/analyze_bimodality.py`). Statistics behind §5.5.1: unclamped, the instruct model is strongly bimodal at its midpoint dose (dip $p < 10^{-4}$), but the bimodality vanishes under the clamp ($p = 0.19$ – 0.46); clamped, the *base* model is the all-or-none one (dip $p \leq 0.013$ pooled, modes at $E \approx 0.28$ and 0.91 , 65–90% of prompts individually significant at threshold doses). Guards: low- k zero-inflation at the binarization threshold is a thresholding artifact and is not counted as bimodality, and per-prompt significant fractions agree with the pooled calls throughout.

Appendix D — Validity gates and correctness checks

All gates are bitwise assertions run inside the jobs, before any statistic is computed.

- **Determinism gate** (every protocol): two identical clean runs must produce bitwise-identical traces before any Δ is trusted.
- **Zero-pulse clamp gate** (TIH and all clamped ISI runs): a KV-clamped row with zero pulse must reproduce the free-running clean trace bitwise — passed on all 100 (model, prompt) pairs in the TIH corpus.
- **KV lobotomy / P7** (Figure 3): masking attention to positions $\leq$ the pulse token collapses clamped influence to ≈ 0 *exactly*. After clamping, the masked KV entries are the only causal route, so P7 holds by construction — it is a correctness check on the clamp implementation, not evidence about the model, and is scored as such in §4.

- **Anesthesia wrapper gates** (Figure 4): both lesion wrappers reproduce the unlesioned model bitwise at zero dose, in both arms, so dose-response curves start from a verified baseline.
- **Bimodality guards**: Appendix C.

Appendix E — Free-running-only results and protocol history

Findings quarantined here are real observations under the *uncorrected* protocol; none is quotable against the clamped numbers in the main text.

- **The pre-registered Δ AIC straw man, corpus-wide.** The pre-registered logistic-vs-linear statistic had called 35 of 36 result cells ignition-like; the nested Hill $n = 1$ baseline keeps 21, with every flip toward graded saturation (`results/isi_hill_baseline.md`). This is the §3.3 correction's motivation and its cost.
- **The Qwen2.5 base-model null artifact.** At sham dose the base model already flips **39% of its own continuation tokens** under near-tie greedy decoding, so its "null" is itself a cascade — inflating the binarization threshold θ and suppressing E everywhere. This is what manufactured the published free-running threshold ratio of 0.029 (an order of magnitude below the clamped 0.323) and the base model's apparent ignition-resistance; it is also the concrete instance behind §7's "the null is not a null."

LLM Usage Statement

We used Claude Fable 5 to brainstorm approaches and help draft sections, with lots of steering and nudging here and there, especially the resource allocation and speedup, and the results to highlight. All results and claims were independently verified. Total GPU spend for every number in this report $\approx$ \$19 in Hugging Face credit.