MEASURING THE STABILITY OF AUTOREGRESSIVE GENERATION: PAIRED LATENT TRAJECTORIES UNDER BENIGN INPUT PERTURBATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Replacing a single word in a factual question can change an instruction-tuned model's answer even when the question's meaning is preserved. Across six open-weight models, 8 to 34 percent of our audit-accepted substitutions change the answer. We introduce the Latent Trajectory Assay (LTA) to measure how such changes propagate through generation. LTA pairs hidden-state trajectories for clean and perturbed prompts in two conditions. Letting the perturbed prompt generate its own continuation shows whether the answer changes. Under teacher forcing, the clean continuation is fixed through both prompts, which isolates the internal response to the input change. In all six models the internal divergence, and how well it predicts a changed answer, grow with depth. Most perturbed trajectories drift back toward their clean counterparts, but so do trajectories under semantically inert edits such as a change of letter case, and what an audited lexical substitution adds is a larger displacement that outlasts that baseline. Excluding the eight-position answer window, divergence still discriminates answer flips at AUROC 0.68 to 0.76 and leads a paired output-probability score from the same pass on every model by point estimate. A human audit of 360 substitutions rejects 14 percent as meaning-changing. The signal survives their removal (AUROC 0.68 to 0.78), and it flags the rejected substitutions on its own (AUROC 0.76). For screening fragile prompts before a rewording is known, black-box resampling beats every hidden-state statistic, and an erosion score from the teacher-forced pass improves that screen on five of six models without hidden-state access. LTA's contribution is therefore not a better detector but a controlled, depth-resolved measurement of the internal response to a specific input change. We release the instrument, capture archive, and tooling.

1 INTRODUCTION

Language-model reliability is usually measured at the output, from confidence or from agreement across resampled generations, while the internal computation that produces failure goes largely unmeasured. We replaced single words of factual questions with audit-accepted intent-preserving substitutes and held decoding fixed. Llama-3.1-8B-Instruct, the model where this investigation began, then changed its answer on more than one in five perturbed prompts, and five other models flipped on between one in twelve and one in three (Table 1). Confidence gave at best an indirect warning about which prompts were fragile and said nothing about where the fragility originated. The black-box posture was once forced on the field, but open-weight models now sit within reach of frontier capability (Sajadieh et al., 2026), and many consequential deployments expose every hidden state to their operators. Interpretability, meanwhile, concentrates on explanation (Gurnee et al., 2026) rather than on behavioral stability.

Autoregression already supplies the object our instrument measures. A generating model emits one internal state per token, and a sequence of states responding to a controllable input is an object the experimental sciences know how to interrogate, by disturbing the input a measured amount, holding everything else fixed, and following the response over time and over depth. A growing body of work

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

treats generation as a dynamical object in this spirit (Cherukuri & Varshney, 2026; Zhang et al., 2026; Sun et al., 2026; Su et al., 2024), but on trajectories produced by uncontrolled generation, where the input, the model's own decoding feedback, and the depth of the readout all arrive entangled. To our knowledge, no design in this line combines a measured input perturbation with a fixed decoding path that separates input-driven state change from a diverging token sequence.

We therefore introduce the Latent Trajectory Assay (LTA), a paired measurement design that compares clean and perturbed generations while separating input-induced state change from divergence caused by autoregressive feedback. It treats one generation as a trajectory of hidden states and compares a clean prompt against a copy in which one word is replaced by a synonym that the language-model audit accepted as intent-preserving. The comparison runs position by position through the generation, at four depths in the network, and in two conditions. In the free-run condition the perturbed prompt generates its own continuation, as in deployment. In the teacher-forced condition we force the clean continuation through the model under the perturbed prompt, so the token paths are identical and any state divergence is attributable to the input alone. The instrument is observational and label-free, and its only inputs are the hidden states that every forward pass already produces.

The study is exploratory, and we ask three questions. The first is how generation trajectories respond to a single measured, intent-preserving substitution, over the course of generation, across depth, and across models (Section 5, with the dose axis and the unperturbed dynamics in Appendices K and E). The second is whether the size of that response predicts which answers will change, without labels, and how it compares against confidence read from the same forward passes (Section 6). The third is what the internal measurement adds over black-box resampling at a matched budget, alone and in combination (Section 6.1). We make three contributions.

- **An instrument.** The Latent Trajectory Assay measures how a generating model responds internally to a controlled change in its input. It fixes the decoding path, pairs hidden states position by position at four depths, and separates the response to the input from the response to a changed continuation. We release it with its capture, analysis, and audit tooling.
- **A phenomenon.** One synonym swap moves the hidden states of all six open-weight models, from 4B to 70B parameters, by an amount that grows with depth and varies widely across prompts (Figure 1). The displacement decays along the continuation about as often as it does under a semantically inert edit, but an audited swap lands farther away and more of it persists (Figure 10).
- **A label-free instability signal, and what validates it.** The size of that displacement predicts which rewordings will change the answer, with no labels (Figure 2), even when the score is denied the positions where the answer appears (Table 13). It beats an output-probability score from the same pass, survives a human audit of the substitutions, and flags the substitutions the auditors reject. For screening prompts before a rewording is known, resampling performs best, and the best screen adds a teacher-forced score to it (Figure 8, Table 8).

2 Related work

Our study connects three lines of work. The first treats autoregressive generation as a dynamical process in representation space, spanning the hallucination basins of Cherukuri & Varshney (2026), DMET (Zhang et al., 2026), the step geometry of Sun et al. (2026), single-generation hallucination detection in MIND (Su et al., 2024), and TILR (Malarkkan et al., 2026). Every trajectory there comes from uncontrolled generation, which entangles input effects, decoding feedback, and depth. Our contribution is experimental control. The ICR Probe (Zhang et al., 2025) tracks a different trajectory, the cross-layer evolution of one state. Professor Forcing (Lamb et al., 2016) showed that hidden-state dynamics under teacher forcing and free-running generation differ, and closed the gap as a training objective. We use the same contrast as a paired observational control.

The second line takes stability under input perturbation as the quantity to measure. Our nearest neighbor is Lyapunov Probes (Luan et al., 2026), which shares that premise, and we differ on the three axes that define the instrument. Our signal is label-free where theirs is supervised on correctness labels, our operator is one controlled meaning-preserving substitution where theirs mixes text edits with hidden-state noise, and our object is the evolving trajectory rather than a static representation.

They impose stability as a training constraint, and we measure it. CCPS (Khanmohammadi et al., 2025) and Liu et al. (2026) perturb internal states, not the input, and SPUQ (Gao et al., 2024) perturbs prompts, not internals. S^2R^2 (Li et al., 2026) shares the structural core of our teacher-forced condition, decoder states compared between clean and perturbed inputs under the same forced reference target. It works at segment-mean granularity and serves as a LoRA robustness loss rather than as a per-position, depth-resolved inference-time measurement. POSIX (Chatterjee et al., 2024) provides the closest precedent for the rescoring primitive behind our erosion score (Section 6.1), averaging absolute log-likelihood shifts over intent-preserving variants into a model-level index. We make it directional, single-swap, per-pair, and per-query. Wen et al. (2026) study an analogous quantity in embedding space, and Xia et al. (2026) argue for evaluating confidence under such variation. Since Sadanandan & Behzadan (2026) already predict paraphrase-induced flips from single-pass entropy, flip predictability is not our contribution and we benchmark against it. Conti et al. (2025) and TriLens (Yang et al., 2026) read depth in a static forward pass, not divergence over depth during generation.

The third line reads reliability from label-free signals at a fixed input. A mature family resamples outputs, from semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024) and its probes (Kossen et al., 2024) through SelfCheckGPT (Manakul et al., 2023), INSIDE (Chen et al., 2024), LLM-Check (Sriramanan et al., 2024), the geometric entropy of Medina et al. (2026), and a Koopman-based variant (Wilson & Akrout, 2026). That family varies the sampling path, while we vary the input at a fixed decoding path. Local geometry is an established reliability signal for single states, where Yin et al. (2024) read local intrinsic dimension and Anonymous (2026) link per-query neighborhood structure to representation mobility under benign input perturbation. We adapt that construction to generation as a control (Section 7).

3 The Latent Trajectory Assay

3.1 Trajectories

We treat one autoregressive generation as a trajectory through the model's latent space (Appendix B gives the design schematic and a notation table). For a prompt of P tokens and a generation of T tokens, the forward computation yields a residual-stream state at every block and position. Writing $h_i^\ell \in \mathbb{R}^d$ for the output of block ℓ at position i of the prompt-plus-generation sequence, generated token t occupies position $P+t$, so the state of generated token t and the trajectory at depth ℓ are

$$h_t^\ell \;\equiv\; h_{P+t}^\ell, \qquad \tau^\ell \;=\; \big(s_1^\ell, s_2^\ell, \ldots, s_T^\ell\big) \in \mathbb{R}^{T\times d}, \qquad \ell(r) \;=\; \max\big(0,\; \lfloor rL \rceil - 1\big), \tag{1}$$

where s_t^ℓ is the state at step t under whichever of the three definitions of Section 3.2 is in use, with h_t as the default, L is the model's block count, and the four relative depths $r \in \{0.25, 0.5, 0.75, 1.0\}$, referred to as early, early-mid, late-mid, and final, fix the recorded block indices ($\lfloor\cdot\rceil$ rounds to the nearest integer). The state h_t has consumed the prompt and the first t tokens of the continuation and conditions token $t+1$. The trajectory τ^ℓ is the object of every analysis. A pair compares two trajectories row by row (Section 3.3), and the per-trajectory dynamics of Appendix E are functions of consecutive rows, the transition magnitude $\|s_{t+1} - s_t\|$ and the meander ratio, path length over net displacement. We also record each generated token's next-token entropy in bits and log-probability, which come from the state at position $P+t-1$ that conditions it, and we respect that alignment wherever states meet token-level quantities.

3.2 Three definitions of the state at step t

Which vector constitutes the state s_t at step t is itself a design decision, so we carry three definitions and compare them empirically. The first is the last-position residual state h_t of Equation 1, the vector that conditions the next-token distribution (the prefix cache, not this vector, is the decoder's full state). The second is the prefix mean m_t and the third the attention-weighted context O_t,

$$m_t \;=\; \frac{1}{P+t}\sum_{i=1}^{P+t} h_i, \qquad O_t \;=\; \sum_j \bar{a}_{t,j}\, x_j, \tag{2}$$

where $\bar{a}_{t,j}$ is the head-averaged attention weight and x_j the block-input states it attends over. Substituting m_t or O_t for s_t in Equation 1 gives the trajectory under that representation. The prefix mean is

a low-pass summary whose sensitivity to each new token decays as $1/t$, and paired comparisons stay valid because a clean and a perturbed trajectory carry the same smoothing. Where h_t is the readout state, O_t is an attention-weighted summary of the context it reads, not the attention block's projected output. Because it requires materialized attention weights and the slower eager attention path, we record it on the three models at or below eight billion parameters.

3.3 Paired divergence measures

The unit of analysis is a pair of trajectories: the clean trajectory (τ^c), and the trajectory of the same prompt under a small input perturbation (τ^p), at one depth and state representation. Writing s_t^c and s_t^p for their states at aligned step t, we compute the per-position divergences

$$d_t^{\cos} \;=\; 1 - \frac{\langle s_t^c,\, s_t^p \rangle}{\|s_t^c\|\,\|s_t^p\|}, \qquad d_t^{\mathrm{euc}} \;=\; \|s_t^c - s_t^p\|_2, \tag{3}$$

and summarize each curve by its mean, maximum, and final value. For free-running pairs we record the divergence onset, the first position where the token paths split. The contraction slope, the least-squares slope of $\log d_t^{\mathrm{euc}}$ over generation steps t, characterizes whether a perturbed trajectory pulls back toward its clean counterpart (negative values) or drifts away (positive values). We compute divergences in double precision on half-precision states, in the raw residual-stream basis, which we verified does not change the conclusions (Appendix J).

4 Experimental design

4.1 Task and data

We evaluate LTA on open-ended factual question answering, using the validation split of TriviaQA in its no-context configuration (Joshi et al., 2017), sampled once with a fixed seed and reused for every model. A fixed instruction asks for a short answer and a one-paragraph justification, which extends generations to a length sufficient for trajectory statistics (median aligned lengths run from 76 to 139 positions across the models). We sample 500 questions and evaluate the 494 for which the audited operator of Section 4.3 yields at least one audit-accepted variant, reduced to 245 for the largest model to fit its compute budget. We grade correctness by normalized match of the extracted short-answer span against the gold alias set, without an external judge.

4.2 Models

Six open-weight instruction-tuned models span three families at two scales each, Llama-3.1-8B-Instruct, Gemma-3-4B and Qwen3-8B, then Gemma-3-27B, Qwen3-32B and Llama-3.3-70B-Instruct (Table 2, Appendix A). The investigation began on Llama-3.1-8B-Instruct, and we use that model for the analyses that run on a single model: the representation comparisons, the dose axis, and the two-arm screening analysis of Section 6.1. All models run in bfloat16 with their native chat templates, and Qwen3 models run with thinking mode disabled.

4.3 Perturbation operator

Perturbations are synonym substitutions in counter-fitted word-vector space (Mrkšić et al., 2016), a standard construction whose meaning preservation is usually assumed rather than checked. We audited 1,500 candidates drawn from that space over the 500 questions and found 44 percent of them not meaning-preserving (interval 0.42 to 0.47, cluster bootstrap over questions), so we gate and audit the capture operator. We filter candidates against an English lexicon by frequency and send six per question to a language-model audit that judges them in context (2,993 gated candidates, 62 percent judged benign, interval 0.60 to 0.64). The first $K = 3$ audit-accepted candidates per question form the variant set, 1,376 audited variants over 494 of the 500 questions. We release them with the capture and illustrate them in Appendix C. The audit is a language-model judgment, not a human annotation. A Claude-family assistant (mid-2026) classified every gated candidate in the context of its question into five categories (good synonym, morphological variant, meaning shift, cross-lingual, junk), with borderline cases resolved to meaning shift, and we accepted the first two categories in candidate order. We call the variants audit-accepted, and reserve verified for the human-confirmed subset that follows.

Two annotators then rated a stratified sample of 360 variants, 90 in each cell of audit category by flip status on any model, blind to both, on a four-level scale (equivalent, answer preserving, intent changed, broken), and we treat equivalent and answer preserving as accepted. The first annotator rated every item and the second a 120-item overlap. Rating independently, the two agreed on the accept decision for 87 percent of the overlap, kappa 0.32 at a rejection base rate of 7 to 14 percent. Half of the 16 disagreements were swaps inside a title or a proper name, a class the protocol's written rule places at intent changed and the second annotator had not applied. On the other 110 items the independent ratings agreed on 93 percent of accept decisions (kappa 0.52). Applying the written rules to both sheets with flip status still hidden raised accept-decision kappa to 0.96, and Appendix D gives the agreement in full and the result under every acceptance rule. Weighted to the audited population, the first annotator rejects 14.1 percent of audit-accepted variants (interval 10.4 to 17.7), 19.1 percent of those that flipped on at least one model against 7.6 percent of those that flipped on none, so about one flipping variant in five is a swap a human rejects. A variant is confirmed when every annotator who saw it accepted it, one annotator for the 240 items outside the overlap. Removing the rejected variants lowers the pooled flip AUROC of the primary score by 0.031 (interval 0.015 to 0.051), from 0.766 to 0.735, with per-model values of 0.68 to 0.78, and the reduction is 0.025 to 0.038 under the alternative acceptance rules (Table 5). The lead over the best confidence baseline holds on four of six models by point estimate and excludes zero on two. Variants rated intent changed flip at about twice the rate of accepted variants and carry nearly three times their median divergence, and the score predicts which variants a human rejects at AUROC 0.71 to 0.79 per model (0.76 pooled, interval 0.69 to 0.82), so the score flags meaning-changing swaps without labels. Ungrammatical but intent-preserving input is in scope by design, since deployed inputs need not be grammatical. Sixty percent of variants are ungrammatical for at least one annotator, and the score shows little association with grammaticality. It separates ungrammatical from grammatical variants at AUROC 0.48 to 0.55, median divergence differs between the two groups by about 0.001 or less in either direction on every model, and flip rates are 0.230 on grammatical against 0.237 on ungrammatical variants. The confirmed-and-grammatical subset (112 items) scores 0.755 with a change of -0.011 whose interval (-0.059 to 0.036) does not settle its sign, and the 63 variants every annotator who saw them rated equivalent score 0.81 at a size that limits precision.

Each variant applies $m = 1$ substitution to the question text, never to the instruction scaffold. On Llama-3.1-8B we additionally run $m \in \{2, 4\}$ substitutions as a graded dose axis (Appendix K). Those arms predate the audit and draw on the raw substitution space at every dose, including their own single-swap reference arm, so they measure response to graded lexical disturbance rather than to audit-accepted variation.

4.4 Conditions

We measure each clean-variant pair under two conditions (Figure 3). In condition (a), free-run, the model generates greedily from the perturbed prompt, so latent-state and token-path differences compound, and the onset records where the token paths split. In condition (b), teacher-forced, we push the clean generation's tokens through the model under the perturbed prompt, so the trajectories share a token path and any state divergence is attributable to the input alone. Condition (b) measures the response to the input change along a common forced continuation, not the perturbed model's own trajectory, which for a flipping pair the model would not produce. Its first position, the state of the shared format token that opens every continuation, carries no answer content, and the curve that follows shows how the displacement present at the prompt boundary evolves along the controlled path. We establish the relevance of this controlled condition empirically: divergence measured under teacher forcing predicts the answer flips observed in free-running generation.

Condition (c) resamples, five generations per prompt at temperature 0.7 from the unperturbed input, on every model for all 494 prompts of each audited capture and the 70B's 245-prompt subset. On Llama-3.1-8B the mean agreement rate is 0.75 and 35 percent of prompts give more than one answer cluster, while across the grid agreement runs from 0.61 on Gemma-3-4B to 0.94 on Llama-3.3-70B. This control separates input perturbation from sampling consistency. Decoding is greedy outside condition (c), and the capture completed every prompt attempted, 494 of 494 on five models and 245 of 245 on Llama-3.3-70B-Instruct, with zero failures.

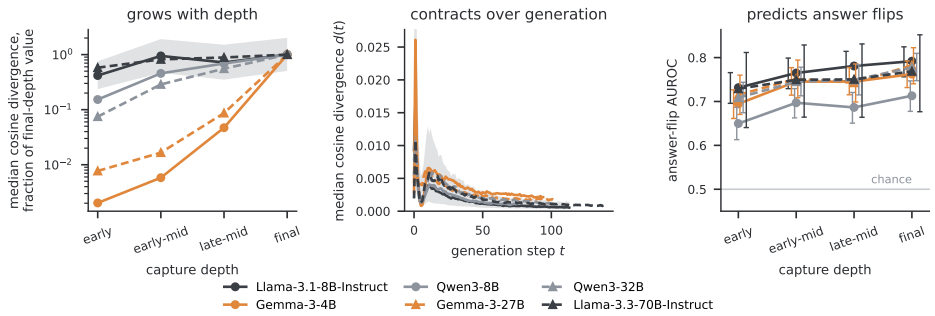

Figure 1: The stability phenomenon across six models. Left: median teacher-forced divergence by depth, normalized to each model's final-depth value (log scale). Cosine divergence grows with depth in every model, at a steepness characteristic of the family, from a factor of about two (Llama) to 130 to 490 (Gemma, whose early-depth cosine is suppressed by massive-activation dimensions). Center: median divergence over generation steps. The perturbed trajectory contracts back toward clean in 92 to 97 percent of pairs in every model (inert perturbations, 95 to 98 percent, Appendix L). Right: answer-flip AUROC of the same divergence by depth, the one depth-resolved quantity that compares across families. It rises from the earliest to the final quarter of the network in all six models (0.65 to 0.73 early, 0.71 to 0.79 final), and the paired gain from earliest to final, read on shared resamples, runs from 0.04 to 0.07 with an interval excluding zero on every model. Shaded bands in the first two panels are interquartile ranges across pairs for Llama-3.1-8B, clipped at the top of the center panel over the opening steps. Bars in the right panel are 95 percent cluster-bootstrap intervals over prompts, offset sideways within each depth. The thin line is chance.

4.5 Utility protocol

We evaluate trajectory-level instability scores (mean, maximum, and final divergence, onset, and contraction slope, per depth and representation) as label-free predictors of answer flips and correctness, measured within our own data. We report AUROC with paired nonparametric tests across prompts and intervals from a cluster bootstrap over prompts (1,000 resamples, 2,000 for the stacked-screen comparisons), scoring fitted components out of fold. We combine scores with a logistic regression over the named features, impute and standardize within each training fold, validate by five-fold cross-validation grouped by prompt, and score by pooled out-of-fold AUROC. Baselines are mean and minimum realized-token log-probability, mean next-token entropy, and a single-state ablation. The sampling-consistency comparison uses the stricter protocol of Section 6.1, which denies every deployable predictor the target pair's own capture.

5 Stability under benign perturbation

We choose the primary representation empirically, using one finding from the atlas of unperturbed greedy generation (Appendix E). On the three models that record all three state definitions, the last-token state h_t diverges under a swap by a factor of roughly 200 to 3,000 more than the attention-weighted context O_t. Free-run divergence saturates once the token paths split, while teacher-forced divergence stays far from saturation and separates stable from fragile prompts by an order of magnitude (Appendix F), so we use condition (b) for the primary divergence analysis.

Every model in the grid shows the same depth-structured, contracting response (Figure 1). Median teacher-forced divergence at the final depth sits between roughly 0.004 and 0.007 in cosine distance for all six models, with a wide distribution across prompts. Within each model the displacement grows from the earliest to the final quarter of the network, at a steepness characteristic of the family. The growth spans a factor of about two in the Llama family, roughly six to thirteen in Qwen, and 130 to 490 in Gemma, where massive-activation dimensions suppress early-depth cosine divergence (Appendix G). Raw magnitudes are therefore not directly comparable across families. A quantity that is comparable is the flip-predictive power of divergence, which rises from the earliest to the final recorded depth in all six models, from AUROC 0.65 to 0.73 to 0.71 to 0.79 (Figure 1, right),

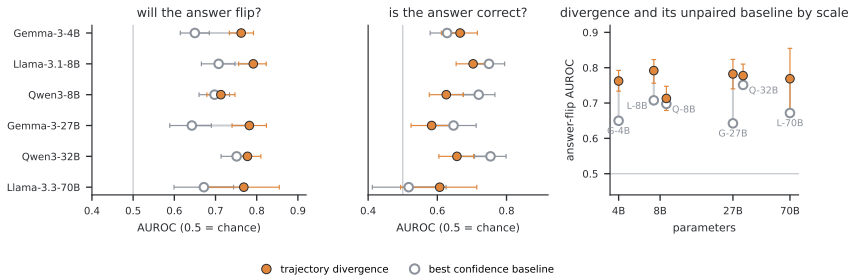

Figure 2: Divergence tracks the flip criterion and confidence tracks correctness. Left and center: per model, AUROC of teacher-forced cosine divergence at the final depth (filled) against the best confidence baseline from the same forward passes (open), on the answer-flip and correctness questions, with 95 percent cluster-bootstrap intervals over prompts. Divergence leads on flips for all six models by point estimate, with overlapping intervals on Qwen3-8B, Qwen3-32B, and Llama-3.3-70B. Confidence leads on correctness for four of six, and the baseline sits at chance (0.52) on Llama-3.3-70B. Right: the flip-question pair against parameter count, intervals on the divergence markers. The thin line is chance (AUROC 0.5). These baselines are unpaired. The paired output-space comparator with the answer window excluded is in Table 13.

with the gain excluding zero on every model. After the initial displacement, in 92 to 97 percent of pairs the perturbed trajectory moves back toward its clean counterpart over the course of generation (per-model values with intervals in Table 1). An inert control perturbation, a scaffold-word swap or a case change, contracts in 95 to 98 percent of pairs, so most of that decay is the shared continuation at work.

The response also scales with the size of the disturbance. On Llama-3.1-8B we widened the perturbation to two and four substitutions, using the raw substitution operator at every dose because a per-swap audit does not establish that a composed perturbation preserves meaning. At every depth, doubling the number of swaps multiplies the median displacement by 2.3 to 2.9 from one swap to two and by 2.2 to 2.4 from two to four, above the factor of two a proportional response would produce, although the raw operator confounds dose with the chance of a meaning-changing swap (Appendix K).

6 Validating the measurement against answer flips

On the flip question, trajectory divergence leads every confidence baseline computed from the same forward passes, on every model. Final AUROC runs from 0.71 on Qwen3-8B to 0.79 on Llama-3.1-8B, against best baselines of 0.64 to 0.75. The margin over the strongest baseline, variant entropy on five models and negative variant log-probability on the 70B, is decisive on Llama-3.1-8B (+0.08, bootstrap 95% interval 0.04 to 0.13), Gemma-3-4B (+0.11, 0.07 to 0.15), and Gemma-3-27B (+0.14, 0.09 to 0.19), marginal on Llama-3.3-70B (+0.10, interval -0.008 to 0.208, on 245 prompts), and absent on the Qwen models (+0.02 and +0.03, both intervals spanning zero). Those baselines are unpaired, and the first eight forced positions, the answer window, are where the score has direct access to the forced clean answer span. With the window excluded from both scores (Table 13, Appendix L), the latent score reads 0.68 to 0.76, a drop of 0.02 to 0.03 with every interval excluding zero, and leads the paired absolute log-odds erosion of the forced tokens over the same positions (0.66 to 0.74) on every model by point estimate, credibly on Llama-3.1-8B, Gemma-3-4B, and Llama-3.3-70B. Section 4.3 gives the same comparison on the human-confirmed subset. We chose the primary score, the mean teacher-forced cosine divergence at the final depth, before running the grid and used it unchanged on every model (robustness across the other summaries in Appendix H).

The correctness question reverses the ordering (Figure 2). Confidence baselines lead correctness prediction on four of six models, and we do not interpret divergence as a general correctness predictor. On the two exceptions, Gemma-3-4B and Llama-3.3-70B, the instability score leads, and on the 70B the best confidence baseline has decayed to chance (0.52) while divergence still reads 0.61 on correctness and 0.77 on flips. Within every family the larger model flips less (0.23 to 0.08 in Llama,

0.34 to 0.13 in Gemma, 0.26 to 0.22 in Qwen), and divergence AUROC remains similar across scales. Its lead over the best baseline is directionally wider in each family's larger model, by 0.01 to 0.03, differences too small to interpret.

We located the signal within the trajectory by ablation. The first teacher-forced position alone reaches AUROC 0.69 to 0.77. The trajectory mean adds a median 0.03 over it, and that gain is the answer window, since with the window excluded the paired difference between the mean and the first position spans zero on five of six models (Appendix L). Much of the predictive signal is therefore already present at the prompt boundary, and what the trajectory uniquely supplies is the depth-resolved relaxation and its timing rather than a better flip score. A paired divergence of the full next-token distributions along the same forced path matches the latent score over the whole trajectory on Llama-3.1-8B (0.80 against 0.79, and stacking the two adds nothing), but the parity comes from the answer window, where the output distribution reads 0.88 because it reads compatibility with the forced answer span almost directly. With the window excluded the state leads, 0.76 against 0.72 (interval on the difference 0.01 to 0.06), and at the prompt boundary it leads by 0.11 (0.75 against 0.64), an advantage observed under the answer-first format, which pins the first output token. A comparison at the first answer-bearing position would sit much closer to the answer event and would no longer test the prompt-boundary distinction (Appendix L). The same appendix shows that prompt-length changes do not carry the signal and that average precision runs 0.40 to 0.61 against base rates of 0.08 to 0.34. On four of six models most flips are changes between two wrong answers, and even so the score separates correct-to-wrong flips from non-flips at AUROC 0.76 to 0.87, ahead of confidence on all six models. The divergence-onset statistic reaches AUROC above 0.93 on every model and we exclude it from all claims, since the answer-first output format makes an early token split and a flipped answer nearly the same event.

6.1 What resampling buys, and what it cannot

The resampling control asks whether five sampled generations per prompt already reveal which prompts are fragile. We score every deployable predictor of the flip at pair level with the target pair's own capture excluded, as an advance fragility screen, on both the audited capture and the earlier raw-operator arm (Table 7, Appendix I).

Black-box dispersion is the stronger advance-screening signal. On the audited arm the answer disagreement rate and the embedded answer-spread score both reach AUROC 0.75, while the best deployable internal statistic, the forced-pair divergence of the same prompt's other variants, reaches 0.70. That deficit of 0.05 has a bootstrap interval (-0.102 to -0.002) excluding zero, and the raw arm reproduces the ordering.

Scored on its own capture, same-pair divergence reaches 0.79, which we use as a non-deployable same-pair reference, against 0.70 for its leave-one-out transfer. That gap reflects information specific to the observed rewording that an advance screen, by design, cannot use. Divergence is informative about this rewording, not merely about the prompt, and flips still occur among prompts whose five resamples agree perfectly, where only the paired measurement separates them (Figure 8, Appendix I). A flip is an output-boundary event, and resampled answers estimate proximity to that boundary relatively directly. A state-space measurement without that information is therefore plausibly disadvantaged on the screening task.

We therefore asked whether reading the paired response through the model's own output head recovers some of that boundary information. We scored each forced pair by the log-odds erosion of its committed continuation, the drop in per-token log-probability of the forced clean tokens between the clean and the perturbed context over the first eight positions, the answer window. We transferred the score across sibling variants exactly as for divergence. The signed form, which we chose before looking at the scores, reaches 0.69 on the audited arm, indistinguishable from transferred divergence. The absolute form, which we evaluated post hoc after the POSIX sensitivity index (Chatterjee et al., 2024) and report as such, reaches 0.75, statistically indistinguishable from the best black-box score and ahead of the signed form on every model in the grid (Appendix I, Table 8). Stacked on the black-box scores, transferred absolute erosion lifts the screen from 0.75 to 0.79 on Llama-3.1-8B, level with the same-pair reference, and the gain replicates on four of the five remaining models. The best deployable screen in this study therefore needs no hidden-state access at all.

7 Discussion

The shared-continuation control shows that contraction should not be read as recovery on its own. Most perturbed trajectories contract toward their clean counterparts, in 92 to 97 percent of pairs, but inert perturbations contract about as often, 95 to 98 percent, so much of the apparent recovery comes from feeding the same continuation through both computations rather than from a return to an attractor. What the audited swap adds is a larger initial displacement and, at matched small initial size, a residual above the null (Appendix L). The instrument measures that response as it propagates without attributing the contraction to a recovery mechanism, a measurement that basin and attractor accounts of generation, built on unperturbed dynamics, cannot separate from the shared continuation (Cherukuri & Varshney, 2026; Zhang et al., 2026).

Much of the response's behavioral significance is already visible at the prompt boundary. The first teacher-forced position alone predicts flips at AUROC 0.69 to 0.77, always within 0.06 of the full trajectory, so fragility is largely set by how the perturbed prompt lands in representation space, and generation propagates a displacement it did not create. Since local neighborhood structure predicts this kind of mobility in classifier embedding spaces (Anonymous, 2026), we asked whether clean-state geometry alone identifies the prompts that will flip, and it does not. Neighborhood features computed for every clean state of Llama-3.1-8B's trajectories against a bank of states from disjoint prompts rank fragile prompts at AUROC 0.58 to 0.63 depending on the flip target (intervals 0.54 to 0.71), below every black-box resampling score (0.70 to 0.80), and adding the geometry to the resampling stack lowers it (dose-arm prompts and prompt-level targets, unlike Section 6.1). Over the whole forced path a divergence of the paired output distributions matches the latent score, but only because of the answer window, where the flip is observed. With that window excluded the state leads on Llama-3.1-8B, and leads the paired log-odds score on every model (Appendix L). Hidden-state access reveals the response across depth and captures prompt-boundary information not yet in the next-token distribution under our answer-first format.

Swaps a human rates as changing intent flip at about twice the rate of accepted swaps and carry nearly three times their median divergence (Section 4.3), so an imperfect operator inflates both the apparent fragility and the instrument's discrimination. Removing those swaps lowers the pooled AUROC by 0.03, and substantial discrimination remains on the human-confirmed subset. Semantic preservation is therefore part of the measurement protocol, not a property that can be assumed from the substitution operator.

Stability and correctness are different reliability questions. Confidence leads at predicting correctness on four of six models while divergence leads at predicting flips on every model, and on Llama-3.3-70B confidence sits at chance for correctness while divergence reads 0.61 there and 0.77 for flips (Section 6). LTA's practical value is the controlled measurement of how one input change propagates, not a general-purpose uncertainty score.

The findings are bounded by a design that studies greedy generation on short English factual questions under a single lexical operator accepted by a language-model audit that a 360-item human annotation finds 14 percent too lenient. We did not repeat the screening comparison under human filtering, we chose the absolute-erosion form after seeing the Llama-3.1-8B results, the dose experiment uses the raw operator, and the inert-perturbation null and the output-distribution comparison ran only on Llama-3.1-8B. Appendix M gives the limits in full.

8 Conclusion

The Latent Trajectory Assay provides a controlled way to measure how a single input change propagates through autoregressive generation. It resolves divergence that is depth-structured, heterogeneous across prompts, and contracting about as often as inert perturbations do, and it yields a label-free signal that predicts answer flips on all six models at AUROC 0.68 to 0.76 with the answer window excluded, leads a paired output-space score on every model by point estimate, and survives a human audit of the operator with a reduction of 0.03. Black-box resampling is the stronger standalone advance screen, and on five of six models the strongest screen stacks it with the teacher-forced erosion score. Its distinctive contribution is the per-event, depth-resolved response that answer-space methods cannot observe.

AI USE STATEMENT

In this work, we used generative AI tools (large language model assistants) to provide feedback on the experimental design and analysis protocol, to implement and refactor analysis code, and to assist in interpreting results. We have not used generative AI tools to generate synthetic datasets, to develop theoretical models, or to formulate or prove mathematical claims, and translation and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools to draft and edit parts of the manuscript for readability, to create and revise figures, to search and summarize related literature, and to format references. We have reviewed all AI-assisted work. Every reported number was recomputed from committed analysis scripts and checked against the archived outputs by the authors, every citation was verified against its publication record, and all generated code was reviewed and tested. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work measures the behavior of publicly released open-weight models on a public question-answering benchmark. The only human data are ratings of question pairs by two annotators, one an author and one a colleague, who judged whether a perturbed question still asks what the clean one asks; no personal data were collected and no participants beyond the annotators were involved. The instrument is observational and does not modify model weights or steer generation. We see no direct path from the results to misuse beyond the general dual-use character of reliability measurement.

REPRODUCIBILITY STATEMENT

Section 4 specifies the task, models, perturbation operator, decoding, conditions, and utility protocol, including seeds, cross-validation grouping, and bootstrap procedure. The appendices give the raw depth profiles and the score-robustness checks behind the main tables. At publication we release, under a permissive license, the capture, analysis, and audit code with its tests, the audited variant set and audit records, the per-pair summary tables of every capture, and the per-step divergence and token tables, so that every reported number regenerates from the released scripts and tables without a GPU. The hidden-state archive (268 GB) will be provided through a dataset host, and the capture itself regenerates from the released variant set and seeds in about eleven GPU-hours on four H200s, up to bfloat16 nondeterminism across hardware.

REFERENCES

Anonymous. Sparse neighborhoods are mobile: A measured link between local structure and representation mobility. *Preprint*, 2026.

Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. POSIX: A prompt sensitivity index for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. arXiv:2410.02185.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *International Conference on Learning Representations*, 2024. arXiv:2402.03744.

Kalyan Cherukuri and Lav R. Varshney. Hallucination basins: A dynamic framework for understanding and controlling LLM hallucinations. *arXiv preprint arXiv:2604.04743*, 2026.

Eddie Conti, Alejandro Astruc, Alvaro Parafita, and Axel Brando. Probing the embedding space of transformers via minimal token perturbations. In *IJCAI 2025 Workshop on Explainable Artificial Intelligence (XAI)*, 2025. arXiv:2506.18011.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. SPUQ: Perturbation-based uncertainty quantification for large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, 2024. arXiv:2403.02509.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, T. Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. Transformer Circuits Thread, 2026. https://transformer-circuits.pub/2026/workspace/index.html.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of ACL*, 2017.

Reza Khanmohammadi, Erfan Miahi, Mehrsa Mardikoraem, Simerjot Kaur, Ivan Brugere, Charese H. Smiley, Kundan Thind, and Mohammad M. Ghassemi. Calibrating LLM confidence by probing perturbed representation stability. In *Proceedings of EMNLP*, 2025. arXiv:2505.21772.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. *arXiv preprint arXiv:2406.15927*, 2024.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023. arXiv:2302.09664.

Alex Lamb, Anirudh Goyal, Ying Zhang, Saizheng Zhang, Aaron Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. In *Advances in Neural Information Processing Systems*, 2016.

Zhuoyun Li, Boxuan Wang, Jinwei Hu, Zhenglin Huang, Qisong He, Xinmiao Huang, Guangliang Cheng, Xiaowei Huang, and Yi Dong. Where do prompt perturbations break generation? a segment-level view of robustness in LoRA-tuned language models. *arXiv preprint arXiv:2605.01605*, 2026.

Deyuan Liu, Zecheng Wang, Zhanyue Qin, Zhiying Tu, Dianhui Chu, and Dianbo Sui. Beyond confidence: The rhythms of reasoning in generative models. In *International Conference on Learning Representations*, 2026. arXiv:2602.10816.

Bozhi Luan, Gen Li, Yalan Qin, Jifeng Guo, Yun Zhou, Faguo Wu, Hongwei Zheng, Wenjun Wu, and Zhaoxin Fan. Lyapunov probes for hallucination detection in large foundation models. *arXiv preprint arXiv:2603.06081*, 2026.

Arun Vignesh Malarkkan, Manan Roy Choudhury, Utkarsh Byahut, Yash Ravindra Charde, Vivek Gupta, and Yanjie Fu. Invariant reasoning directions in latent trajectories of language models. *arXiv preprint arXiv:2606.29164*, 2026.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of EMNLP*, 2023. arXiv:2303.08896.

Johanne Medina, Tianyi Zhou, Keivin Isufaj, Aristides Gionis, and Sanjay Chawla. Integrating local and global entropy for uncertainty quantification in LLMs. *arXiv preprint arXiv:2606.09875*, 2026.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. Counter-fitting word vectors to linguistic constraints. In *Proceedings of NAACL-HLT*, 2016.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018.

Binesh Sadanandan and Vahid Behzadan. Predictive entropy links calibration and paraphrase sensitivity in medical vision-language models. *arXiv preprint arXiv:2604.08941*, 2026.

Sha Sajadieh, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Lapo Santarlasci, Juan Pava, Nestor Maslej, Russ Altman, Erik Brynjolfsson, Carla Brodley, Jack Clark, Virginia Dignum, Vipin Kumar, James Landay, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Elham Tabassi, Russell Wald, Toby Walsh, and Dan Weld. The AI index 2026 annual report. Technical report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2026.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. LLM-Check: Investigating detection of hallucinations in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 34188–34216, 2024. OpenReview: LYx4w3CAgy.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. Unsupervised real-time hallucination detection based on the internal states of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14379–14391, 2024. arXiv:2403.06448.

Lihao Sun, Hang Dong, Bo Qiao, Qingwei Lin, Dongmei Zhang, and Saravan Rajmohan. LLM reasoning as trajectories: Step-specific representation geometry and correctness signals. In *Proceedings of ACL*, pp. 26872–26887, 2026. arXiv:2604.05655.

Qihao Wen, Jiahao Wang, Yang Nan, Pengfei He, Ravi Tandon, and Han Xu. Embedding perturbation may better reflect intermediate-step uncertainty in LLM reasoning. *arXiv preprint arXiv:2602.02427*, 2026.

Dan Wilson and Mohamed Akrout. Low-cost black-box detection of LLM hallucinations via dynamical system prediction. *arXiv preprint arXiv:2605.05134*, 2026.

Yuxi Xia, Dennis Ulmer, Terra Blevins, Yihong Liu, Hinrich Schütze, and Benjamin Roth. Calibration is not enough: Evaluating confidence estimation under language variations. *arXiv preprint arXiv:2601.08064*, 2026.

Bohan Yang, Yijun Gong, Zhi Zhang, Ge Zhang, Wenpeng Xing, and Meng Han. TriLens: Per-layer logit-lens entropy for white-box hallucination detection. *arXiv preprint arXiv:2606.01033*, 2026.

Fan Yin, Jayanth Srinivasa, and Kai-Wei Chang. Characterizing truthfulness in large language model generations with local intrinsic dimension. In *International Conference on Machine Learning*, 2024. arXiv:2402.18048.

Yukun Zhang, Qi Dong, and Mengkang Li. Latent trajectory dynamics in large language models: A manifold evolution framework with empirical validation. *arXiv preprint arXiv:2505.20340*, 2026.

Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. ICR probe: Tracking hidden state dynamics for reliable hallucination detection in LLMs. In *Proceedings of ACL*, 2025. arXiv:2507.16488.

A Model grid and compute

Table 1 gives the per-model headline quantities of Sections 5 and 6 with their intervals.

Model	Prompts	Flip rate	Median d (10^{-3})	Depth ratio	Contracting	Flip AUROC
Llama-3.1-8B	494	0.23 [0.20, 0.26]	3.8 [3.5, 4.2]	2.4 [2.2, 2.6]	0.94 [0.92, 0.96]	0.792 [0.756, 0.823]
Gemma-3-4B	494	0.34 [0.31, 0.38]	6.5 [6.0, 7.1]	494.1 [465.3, 530.4]	0.95 [0.94, 0.97]	0.762 [0.734, 0.792]
Qwen3-8B	494	0.26 [0.23, 0.29]	4.7 [4.4, 5.1]	6.5 [6.2, 7.0]	0.92 [0.90, 0.94]	0.713 [0.679, 0.747]
Gemma-3-27B	494	0.13 [0.11, 0.15]	5.8 [5.3, 6.2]	129.9 [120.4, 139.2]	0.97 [0.96, 0.98]	0.782 [0.740, 0.824]
Qwen3-32B	494	0.22 [0.19, 0.25]	6.3 [5.8, 6.8]	13.4 [12.4, 14.6]	0.92 [0.90, 0.94]	0.778 [0.743, 0.810]
Llama-3.3-70B	245	0.08 [0.06, 0.11]	5.2 [4.6, 6.0]	1.7 [1.6, 1.9]	0.95 [0.93, 0.97]	0.769 [0.677, 0.855]

Table 1: Per-model headline quantities with 95 percent cluster-bootstrap intervals over prompts. Flip rate is the share of audited pairs whose answer changes. Median d is the median teacher-forced cosine divergence of the last-token state at the final depth. Depth ratio is the final-depth median over the earliest-depth median. Contracting is the share of pairs with a negative whole-trajectory contraction slope at the final depth. Flip AUROC is the primary score's discrimination of the flip target.

We capture at batch size one, data-parallel over prompts, on four H200 GPUs of a shared eight-GPU node (Llama-3.3-70B shards each worker across an NVLink-connected pair, and the earliest captures ran on V100 nodes). The audited six-model capture takes about eleven hours end to end at that width, from 44 minutes for Qwen3-8B to four hours for Llama-3.3-70B. We store states as half-precision

Model	Blocks L	Hidden d	Depth indices	O_t	Thinking
Llama-3.1-8B-Instruct	32	4096	7 / 15 / 23 / 31	yes	n/a
Gemma-3-4B	34	2560	7 / 16 / 25 / 33	yes	n/a
Qwen3-8B	36	4096	8 / 17 / 26 / 35	yes	off
Gemma-3-27B	62	5376	15 / 30 / 45 / 61	no	n/a
Qwen3-32B	64	5120	15 / 31 / 47 / 63	no	off
Llama-3.3-70B-Instruct	80	8192	19 / 39 / 59 / 79	no	n/a

Table 2: The model grid. Depth indices are the four recorded block indices (early / early-mid / late-mid / final) resolved by Equation 1 from each model's block count. The O_t column marks the models whose attention-weighted context is recorded, and the thinking column marks the hybrid-reasoning models run with their reasoning mode disabled.

sidecar arrays and derive every table from them, so each analysis in this paper below the capture itself is CPU work, reproducible from the archive without a GPU. The audited six-model capture occupies about 120 GB of sidecar arrays (17 to 27 GB per model), the resampling arms about 70 GB, and the raw single-swap and dose arms on Llama-3.1-8B about 80 GB, for an archive of 268 GB in total.

B THE PAIRED DESIGN

Figure 3 shows the paired design end to end, from the audited swap through the two capture conditions to the three readout families the measurement yields, and Table 3 collects the notation of Section 3 in one place.

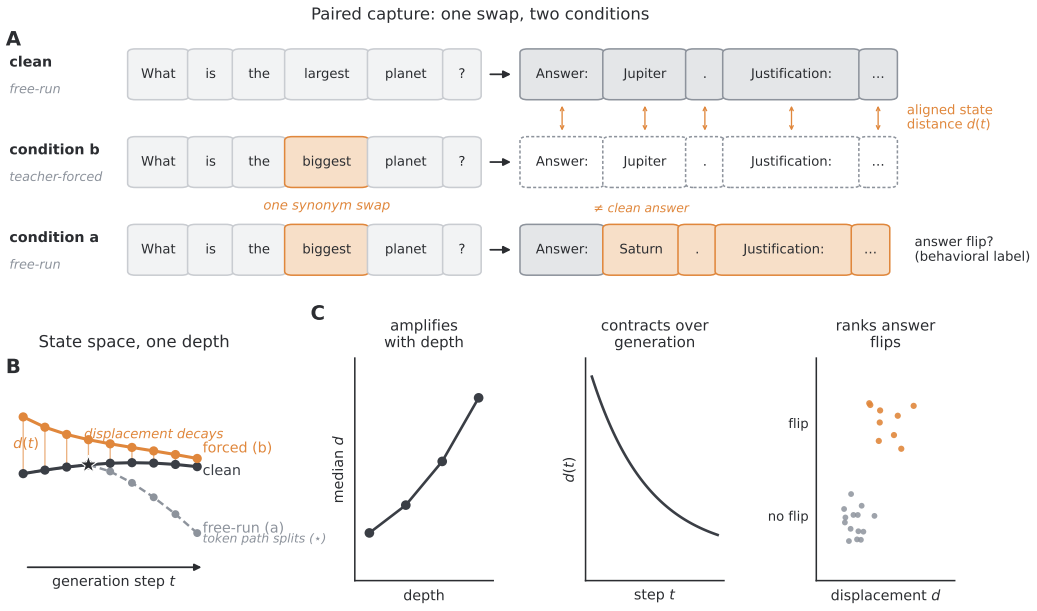

Figure 3: The paired design. (A) A seeded counter-fitted synonym swap produces a perturbed prompt. Under condition (b) the clean continuation is teacher-forced, so states align position by position and the distance $d(t)$ reflects the input alone. Under condition (a) the perturbed prompt generates freely and its answer determines the flip label without external annotation. (B) The design in state space. (C) The three readout families. Panels are illustrative, with measured counterparts in Sections 5 and 6.

Symbol	Where	Meaning
P, T, L	§3	Prompt length in tokens, generation length in tokens, and block count of the model.
h_i^ℓ	Eq. 1	Residual-stream output of block ℓ at sequence position i (prompt and generation together).
h_t	Eq. 1	Last-position state of generated token t, the state at position $P+t$. The default state definition.
m_t, O_t	Eq. 2	Prefix mean of all states up to position $P+t$, and head-averaged attention-weighted context at that position. The two alternative state definitions.
s_t	Eq. 1	The state at step t under whichever definition is in use (h_t, m_t, or O_t).
τ^ℓ	Eq. 1	The trajectory at depth ℓ, the $T \times d$ array whose row t is s_t. The object every analysis operates on.
$\ell(r)$	Eq. 1	Block index recorded at relative depth $r \in \{0.25, 0.5, 0.75, 1.0\}$, called early, early-mid, late-mid, and final.
(τ^c, τ^p)	§3.3	A pair, the clean trajectory and the trajectory of the same prompt under one substitution.
$d_t^{\cos}$, d_t^{euc}	Eq. 3	Per-position divergence between the rows of a pair, cosine distance and Euclidean distance.
Primary score	§6.1	Mean of $d_t^{\cos}$ over the teacher-forced continuation at the final depth, with the answer window excluded where stated.
Answer window	§6	The first eight forced positions, where the clean answer sits and where the score has direct access to the forced answer span.
Onset	§3.3	In a free-run pair, the first position at which the two token paths differ.
Contraction slope	§3.3	Least-squares slope of $\log d_t^{\text{euc}}$ over steps. Negative means the perturbed trajectory pulls back toward the clean one.
Transition magnitude	§3	$\|s_{t+1} - s_t\|$, the size of one step along a single trajectory.
Meander ratio	§3	Path length of a trajectory divided by its net displacement from first to last row.
d_0, residual fraction	App. L	Mean divergence over the first three steps, and mean divergence over the last ten of sixty steps divided by d_0.
Inert perturbation	App. L	An edit that changes tokens but not meaning, such as a case flip, used as the null against audited swaps.
Flip	§4	The free-run answer under the perturbed prompt differs from the clean answer after normalization.

Table 3: Notation of the instrument, with the equation or section that introduces each term.

C AUDITED SWAP EXAMPLES

Figure 4 shows two audited swaps and their decoded outcomes.

(a) Gemma-3-27B. One of two human-confirmed swaps flips the answer, and the justification relocates the evidence.

where are the administrative hq of derbyshire?	*where are the **managerial** hq of derbyshire?*
Answer: Matlock. Justification: Derbyshire County Council's main administrative headquarters are located in Matlock, specifically at County Hall. [. . .]	**Answer: Derby.** Justification: Derbyshire County Council, which serves as the managerial headquarters for the county, is located in Derby city. [. . .] This is reflected in the location of County Hall, the main administrative building, within the city of Derby.

The sibling swap of the same question (*administrative* to *bureaucratic*) answers County Hall, Matlock.

(b) Qwen3-32B. A human-confirmed swap flips a correct answer, and the sibling swaps move it too.

in the british military, a bombadier is equivalent to which other rank?	*in the british military, a bombadier is **comparable** to which other rank?*
Answer: Corporal. Justification: In the British military, a Bombadier is a rank found in the Royal Artillery and is equivalent to a Corporal in other branches of the armed forces. [. . .]	**Answer: Warrant Officer.** Justification: In the British military, a Bombadier is a rank specific to the Royal Artillery and is considered the entry-level warrant officer rank. It is equivalent in rank and status to a Warrant Officer Class 2 in other branches of the British Army. [. . .]

The two sibling swaps (*military* to *soldiers*, *military* to *servicemen*) answer "Corporal or Sergeant" and "Warrant Officer Class 2".

Figure 4: Human-confirmed single substitutions and their decoding outcomes (greedy decoding, identical instruction scaffold, outputs verbatim). Both showcased swaps were rated equivalent in the human audit of Section 4.3. In (a), the swap moves a correct answer to a wrong one and the justification names the right council, then places its County Hall in the wrong town, so the justification conceals the error rather than exposing it, while the sibling swap is inert. In (b), the swap moves a correct answer to a confidently justified wrong one, and the two sibling swaps move it as well, the prompt-level fragility that the resampling screen of Section 6.1 is built to catch. Both models answer correctly on the clean prompt.

D HUMAN AUDIT OF THE OPERATOR

Table 4 gives the per-model values behind the human audit of Section 4.3, and Table 5 the pooled result under every acceptance rule a reader might prefer. Rating independently, the annotators agreed exactly on 67 percent of the 120-item overlap, with quadratic-weighted kappa 0.46 and kappa 0.32 on the accept decision. Sixteen items crossed the accept line, 12 of them accepted by the second annotator and rejected by the first, 8 of those swaps inside a title or a proper name, which the protocol's names-and-titles rule places at intent changed. Kappa on the accept decision is deflated by the rejection base rate, 14 percent for the first annotator and 7.5 percent for the second. Raw agreement was 87 percent, and the prevalence-adjusted kappa is 0.73. The 8 title and name disagreements are one rule, present in the written protocol, that the second annotator had not applied and endorsed on review. On the 110 overlap items outside that class the independent ratings agree on 93 percent of accept decisions (kappa 0.52, prevalence-adjusted 0.86), so the independent disagreement was concentrated in one identifiable class rather than spread across the scale. The reconciliation applied the written rules of the protocol to both sheets with flip status and category hidden, changed seven ratings on the first sheet and fifteen on the second, and raised exact agreement to 79 percent, weighted kappa to 0.70, and accept-decision kappa to 0.96, with one item still across the line. That is adjudicated agreement rather than independent reliability, and the sensitivity table is why the distinction does not reach the conclusions. Against 0.766 on the full sample, the confirmed subset scores 0.735 under the reconciled ratings, 0.741 under the ratings as returned, 0.738 with the 16 disputed items removed, and 0.741 on the overlap items both annotators accepted before reconciliation, and the false-accept rate moves from 15.3 to 14.1 percent. The second annotator rated a 120-item overlap because the first annotator's pace made a full second pass impossible before submission, so confirmation rests

on one annotator for 240 of the 360 items. The score's relation to human rejection is not carried by the flips alone. Among pairs whose answer did not flip on the scoring model it predicts rejection at AUROC 0.70 pooled (interval 0.63 to 0.77), and among variants that flipped on no model at 0.66 (0.55 to 0.75), design-weighted to the audited population like every estimate here.

Model	n sample / confirmed	AUROC, sample	AUROC, confirmed	Confirmed minus sample	Divergence minus baseline, confirmed
Llama-3.1-8B	360 / 313	0.823 [0.764, 0.874]	0.783 [0.715, 0.846]	-0.040 [-0.076, -0.013]	+0.176 [+0.080, +0.269]
Gemma-3-4B	360 / 313	0.784 [0.734, 0.835]	0.782 [0.723, 0.837]	-0.002 [-0.027, +0.022]	+0.095 [+0.011, +0.180]
Qwen3-8B	360 / 313	0.727 [0.666, 0.786]	0.700 [0.634, 0.767]	-0.027 [-0.062, +0.002]	-0.012 [-0.118, +0.083]
Gemma-3-27B	360 / 313	0.770 [0.688, 0.840]	0.736 [0.642, 0.819]	-0.034 [-0.078, -0.001]	+0.099 [-0.020, +0.221]
Qwen3-32B	360 / 313	0.775 [0.715, 0.831]	0.721 [0.652, 0.791]	-0.054 [-0.092, -0.022]	-0.010 [-0.117, +0.108]
Llama-3.3-70B	176 / 153	0.753 [0.523, 0.923]	0.681 [0.388, 0.932]	-0.072 [-0.256, +0.029]	+0.116 [-0.114, +0.374]

Table 4: Flip AUROC of the primary score on the human-audit sample and on its confirmed subset, per model, with the paired change and the paired margin over the best confidence baseline on the confirmed subset. Intervals are 95 percent prompt-cluster bootstraps.

Acceptance rule	Pairs	Flips	Flip rate	Flip AUROC	Change
Confirmed, reconciled ratings	1718	305	0.206	0.735 [0.699, 0.768]	-0.031 [-0.051, -0.015]
Confirmed, ratings as returned	1679	297	0.205	0.741 [0.706, 0.776]	-0.025 [-0.047, -0.004]
Confirmed, disputed items removed	1685	299	0.206	0.738 [0.701, 0.773]	-0.028 [-0.051, -0.008]
Both annotators accept, ratings as returned (overlap)	542	86	0.188	0.741 [0.674, 0.807]	-0.038 [-0.095, +0.005]
Confirmed and grammatical	620	117	0.218	0.755 [0.702, 0.805]	-0.011 [-0.059, +0.036]
Strict (rated equivalent)	348	52	0.183	0.810 [0.736, 0.872]	+0.044 [-0.028, +0.107]

Table 5: The pooled human-audit result under each acceptance rule, six models pooled by within-model percentile rank and design-weighted to the audited population. Pairs and flips are unweighted counts, and flip rates and AUROCs are design-weighted estimates. The change is paired against the full 360-item sample, or against the 120-item overlap for the rule that needs both annotators. Intervals are 95 percent prompt-cluster bootstraps.

E A DYNAMICS ATLAS

Before measuring how generation responds to perturbation, we chart what unperturbed greedy generation looks like through the instrument. First, a trajectory travels remarkably little. Median meander ratios at the final depth run from 52 on Qwen3-8B (interquartile range 45 to 61) to 113 on Qwen3-32B (98 to 128), against median trajectory lengths of 76 to 139 steps, and on Qwen3-32B the median ratio exceeds the median step count outright. The four remaining models fall between, at 61, 73, 84, and 110, with interquartile ranges no wider than 28. Ratios of this size mean the last-token state ends a generation roughly one typical transition away from where it began. A random walk of steps this size would give ratios near the square root of the step count, about ten here, while independent draws from a bounded stationary region give ratios near the step count itself, which is what we observe. Seen from the residual stream, greedy generation samples a bounded region of state space rather than moving through it, as last-token states dominated by token identity would.

The second entry couples the per-step scalars to the state's motion. Writing H_t for the next-token entropy, H_t^{attn} for the attention entropy, and $\|\Delta h_t\|$ for the transition magnitude, we correlate each pair of series over the steps of a single prompt's generation (Spearman, at least thirty aligned steps, the first step dropped) and report the median correlation across prompts. Pooling steps across prompts would manufacture correlation out of prompt-level offsets, so no correlation in this paper is computed that way. Measured within prompts, entropy and transition magnitude are negatively coupled at the final depth in all six models, with medians from -0.12 to -0.32 and interquartile intervals entirely below zero (Figure 5, left). Steps where the model is uncertain are steps where its state moves less, not more. The coupling is also specific to the top of the network. At the earliest recorded depth it is weakly positive in all six models ($+0.01$ to $+0.18$), mixed and mostly small through the middle depths, and uniformly negative only at the final block, so the sign of the relationship between uncertainty and motion is itself a depth-resolved quantity. The early-depth medians are weak enough that the interquartile range straddles zero on Llama-3.1-8B-Instruct (-0.02 to $+0.14$) and Qwen3-32B (-0.07 to $+0.11$), against ranges entirely above zero on the other four, so the early sign is a shift away from the final-depth negative coupling rather than a positive coupling of its own. The

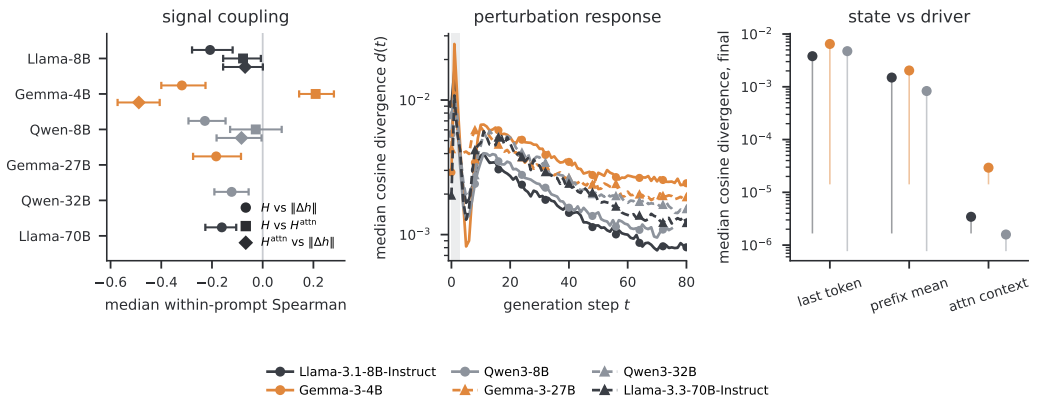

Figure 5: The dynamics atlas across the grid. Left: within-prompt Spearman correlation of the per-step signals at the final depth (median and interquartile range across prompts; 493 to 494 prompts per model, 245 for Llama-3.3-70B). Next-token entropy and transition magnitude are negatively coupled in every model (medians -0.32 to -0.12, interquartile intervals entirely below zero). The attention pairs, recorded on three models, are heterogeneous, and Gemma-3-4B is the outlier ($+0.21$ with entropy, -0.49 with transition magnitude). Center: median teacher-forced cosine divergence over generation steps at the final depth. One audited swap produces an early spike (shaded, 2.0 to 5.8 times the level of steps 20 to 60) and a secondary bump at steps 9 to 13 in every model. Right: the same divergence by state representation on the three models that record all three. The attention-weighted context moves 200 to 3,000 times less than the last-token state (bootstrap intervals 214 to 3,159 over the three models), with the prefix mean 2.5 to 5.7 times below the last token (intervals 2.4 to 6.1).

attention pairs, available on the three models that record O_t, show no comparable regularity. Both couplings sit near zero on Llama-3.1-8B-Instruct and Qwen3-8B, while Gemma-3-4B stands apart, with attention entropy positively coupled to next-token entropy ($+0.21$) and negatively coupled to transition magnitude (-0.49), and we accordingly report the attention couplings as family-dependent rather than universal.

The third entry is the shape of the response to perturbation itself. At the final depth, the median teacher-forced divergence curve has the same waveform in every model of the grid. A single audited swap produces a sharp spike within the first three generated steps, at 2.0 to 5.8 times the steady level of steps 20 through 60, then a fast partial relaxation and a secondary bump at steps 9 to 13, landing at step 11 in four of the six models. The waveform reproduces across three families and parameter counts from four billion to seventy billion. The bump is the template handoff into the justification rather than an intrinsic step count. The format boundary sits at median step 5 to 7, the per-prompt divergence peak trails it by a median of 5 to 6 steps, and realigning the curves on the boundary more than doubles the peak-over-trough prominence (Appendix L). Whatever mechanism produces the bump, the response that Section 5 summarizes is a stereotyped impulse response rather than model-specific noise, which is what licenses comparing its summary statistics across the grid.

The last entry settles the representation question that Section 3 deliberately left open. On the three models that record all three state definitions, the teacher-forced divergence of the last-token state h_t at the final depth exceeds that of the attention-weighted context O_t by a factor of roughly 200 on Gemma-3-4B, 1,100 on Llama-3.1-8B-Instruct, and 3,000 on Qwen3-8B, with the prefix mean m_t between the two at 2.5 to 5.7 times below h_t. Under a cluster bootstrap over prompts the three ratios are 221 (interval 214 to 231), 1,105 (1,064 to 1,151), and 2,982 (2,834 to 3,159), and the prefix-mean ratios 3.2 (3.0 to 3.4), 2.5 (2.4 to 2.7), and 5.7 (5.4 to 6.1), so the three orders of magnitude separating the two representations are not an artifact of the median. The attention-weighted context summary is almost unchanged by the swap while the last-position state diverges hundreds to thousands of times more, so the phenomenon lives in the readout state rather than in the context it summarizes, a comparison of two constructed representations rather than of a state against the block's actual

attention output. Together with the mechanical smoothing of the prefix mean, this fixes h_t as the primary representation for the rest of the paper, a choice made empirically.

F What the instrument sees

Figure 6 shows the instrument's view of two prompts of Llama-3.1-8B, one whose answer survives all three audited swaps and one where all three flip. Flip-prediction AUROC by depth for every model of the grid is the right panel of Figure 1.

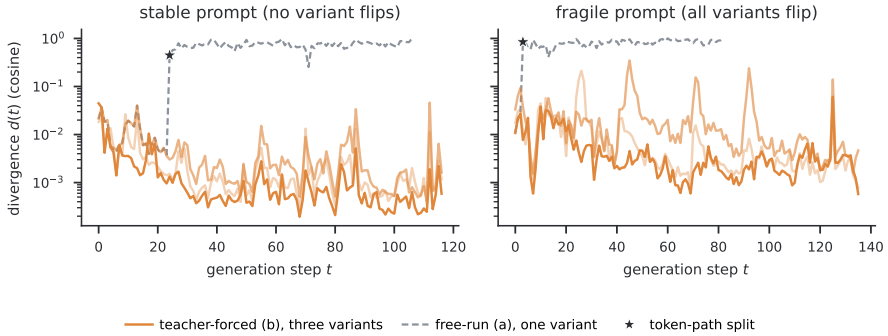

Figure 6: What the instrument sees on two prompts (Llama-3.1-8B, final depth; three teacher-forced variants and one free-run exemplar per prompt; measured data). Teacher-forced divergence (solid) stays orders of magnitude below saturation and separates the two regimes, roughly 10^{-3} on a prompt whose answer survives all three swaps, roughly 10^{-2} on a prompt where all three flip. Free-run divergence (dashed) saturates once token paths split (star), on the stable prompt as well as the fragile one. Token-path divergence is not answer change, which is why we use condition (b) for the stability measurement. Examples are the median members of their regimes, not extremes.

G Raw magnitude depth profiles

Figure 7 gives the median per-pair divergence by depth under both metrics for every model. The two metrics disagree on the Gemma family, which is why the main text reports within-model shapes and label-based depth evidence rather than raw magnitudes compared across families.

H Score robustness

Table 6 reports flip AUROC for every divergence summary against the best confidence baseline, per model. The primary score (mean teacher-forced cosine divergence at the final depth) was fixed before the grid ran and is never swapped per model. On every model it is the best divergence summary or within 0.005 of it, so no comparison in the main text depends on the choice of summary.

model	mean cosine	max cosine	mean Euclidean	first state	last state	best baseline
Llama-3.1-8B-Instruct	0.792	0.760	0.785	0.752	0.733	0.708
Gemma-3-4B	0.762	0.748	0.754	0.707	0.684	0.650
Qwen3-8B	0.713	0.679	0.718	0.691	0.622	0.698
Gemma-3-27B	0.782	0.761	0.783	0.772	0.703	0.643
Qwen3-32B	0.778	0.737	0.780	0.753	0.682	0.751
Llama-3.3-70B-Instruct	0.769	0.766	0.758	0.711	0.747	0.672

Table 6: Flip AUROC by divergence summary (condition (b), last-token states, final depth) against the best confidence baseline from the same capture. The primary score is the first column.

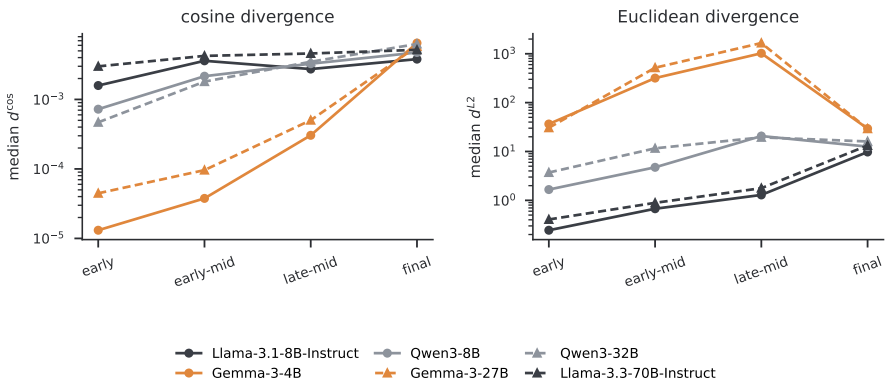

Figure 7: Median per-pair divergence by depth under both metrics (condition b, last-token states, log scale). The two metrics tell opposite stories on the Gemma family, where massive-activation dimensions dominate the direction of the residual stream, suppressing early and mid-depth cosine divergence toward zero, while the absolute motion of those same dimensions inflates the Euclidean distance at exactly those depths. Neither raw magnitude is comparable across families, so the main text reports within-model shapes and label-based depth evidence.

I Screening analysis details

Table 7 gives the full advance-screening comparison on Llama-3.1-8B on both operator arms, and Figure 8 places every forced pair on the two instability axes. The remainder of this appendix records the accounting behind the stacking and erosion results of Section 6.1.

			flip AUROC	
class	predictor	deployable	raw arm	audited arm
black-box	answer spread vs clean (MiniLM)	yes	0.713	0.746 [0.706, 0.782]
black-box	answer disagreement rate	yes	0.702	0.752 [0.712, 0.789]
black-box	answer spread, pairwise (MiniLM)	yes	0.682	0.723 [0.682, 0.765]
black-box	answer cluster entropy	yes	0.664	0.724 [0.685, 0.763]
black-box	mean token entropy	yes	0.622	0.666 [0.622, 0.705]
black-box	mean log-probability (neg.)	yes	0.611	0.659 [0.615, 0.701]
white-box resample	state spread, early-mid	yes	0.601	0.608 [0.559, 0.652]
white-box resample	state spread, early	yes	0.592	0.592 [0.540, 0.638]
white-box resample	state spread, late-mid	yes	0.582	0.597 [0.545, 0.643]
white-box resample	state spread, final	yes	0.573	0.565 [0.516, 0.613]
white-box cross-variant	cross-variant divergence	yes	0.615	0.703 [0.657, 0.744]
same-pair reference	same-pair divergence	no	0.788	0.792 [0.756, 0.823]

Table 7: Advance flip prediction on Llama-3.1-8B, pair level, with every deployable predictor denied the target pair's own capture. Audited-arm values carry 95 percent cluster-bootstrap intervals over prompts; the raw arm's are in the released tables. The audited arm is the capture of record; the raw arm replicates the analysis on the pre-audit operator and that wave's earlier flip-label matcher (flip rate 0.35). Black-box scores derive from the five resampled generations of condition (c), white-box resample scores from the hidden states of those same generations, and cross-variant divergence transfers the forced-pair measurement from the prompt's other variants. The same-pair reference scores each pair from its own capture and is not deployable as an advance screen.

Every stack is built and validated per the protocol of Section 4.5, and the black-box stack combines the six black-box scores of Table 7. Adding resample-derived state features to that stack moves it by about 0.002, and trajectory-shape functionals of the divergence curve add nothing over its mean even with the target pair instrumented. On the confident half of the dose-arm corpus (223 prompts

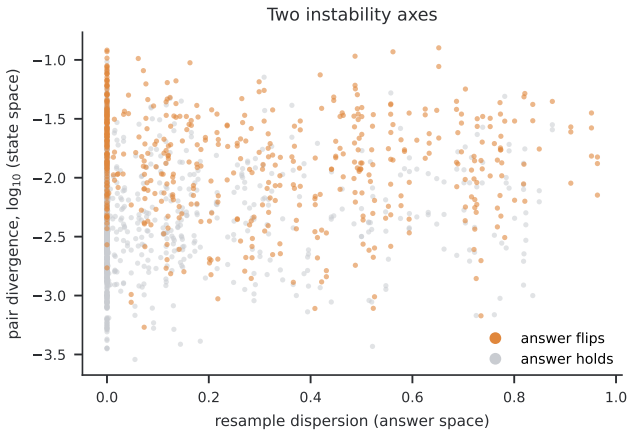

Figure 8: The two instability axes on Llama-3.1-8B (audited arm). Each point is one forced pair, placed by its own teacher-forced divergence (vertical, log scale) and its prompt's resample dispersion, the embedded spread of the five sampled answers against the clean answer (horizontal), with flipped pairs in orange. Flips concentrate toward the top of every vertical band, including the leftmost one, where resampling is perfectly consistent and the answer-space screen is silent.

below the median clean entropy) clean entropy drops to AUROC 0.53 and 0.56 on the two flip targets, intervals that include chance, while the resample stack keeps 0.68 and 0.73, intervals that do not.

On Llama-3.1-8B, absolute erosion transfers far better than the signed form on the audited arm (0.75 against 0.69, interval on the difference 0.03 to 0.08), beats transferred divergence there (interval 0.01 to 0.08), and is statistically indistinguishable from the best black-box score (difference -0.003, interval -0.05 to $+0.04$), although on the raw arm it stays below the black-box screen and its edge over the signed form does not clear zero. Added to the black-box stack, transferred erosion lifts the screen from 0.75 to 0.79 (interval on the gain 0.02 to 0.08, against 0.03 for transferred divergence), level with the 0.79 same-pair reference, in either form.

		standalone screen		stacked screen		
model	flip rate	best black-box	abs. erosion	black-box	+ erosion	gain
Llama-3.1-8B	0.225	0.752	0.749	0.747	0.792	$+0.044$ [$+0.016$, $+0.083$]
Gemma-3-4B	0.344	0.714	0.743	0.714	0.762	$+0.048$ [$+0.023$, $+0.077$]
Qwen3-8B	0.260	0.731	0.711	0.737	0.767	$+0.030$ [$+0.006$, $+0.057$]
Gemma-3-27B	0.131	0.655	0.701	0.621	0.675	$+0.054$ [$+0.002$, $+0.104$]
Qwen3-32B	0.222	0.749	0.774	0.764	0.802	$+0.038$ [$+0.011$, $+0.074$]
Llama-3.3-70B	0.083	0.700	0.713	0.685	0.650	-0.035 [-0.072, $+0.025$]

Table 8: The screening analysis per model, audited captures, pair level, every deployable predictor denied the target pair's own capture. The best black-box column takes the strongest of the six dispersion scores of Table 7 per model. Absolute erosion is the transferred sibling mean of the first-eight-position absolute log-odds shift, and the stacked columns give the pooled out-of-fold AUROC of the dispersion stack without and with it. Gains carry 95 percent cluster-bootstrap intervals over prompts with both stacks refit in every resample, and five of the six exclude zero. Bootstrap intervals on each model's erosion-versus-dispersion difference all span zero.

The accounting generalizes to the five remaining models (Table 8). The absolute form transfers better than the signed form on every one of them, by margins from 0.03 to 0.13 whose intervals all exclude zero, so the observation made post hoc on Llama-3.1-8B is confirmed out of sample. Against each model's own best dispersion score, transferred absolute erosion is statistically indistinguishable everywhere and ahead by point estimate on four of the five. Stacking replicates on four of the five, with gains of 0.03 to 0.05 that bracket the 0.04 of Llama-3.1-8B and at least match the divergence add-on in every case. Llama-3.3-70B is the exception, with no add-on improving its stack, and since

its 677 pairs from 245 prompts hold only 56 flips, the least evidence in the grid, we report the stacking elevation as a five-of-six result.

The signed and absolute erosion forms carry different information. The sign is strong event-level evidence, because the thirty percent of audited pairs whose committed answer becomes more likely under the swap flip at four percent against thirty percent for eroding pairs, yet it cancels when averaged across siblings. The magnitude of the shift, in contrast, is the prompt-level sensitivity that transfers. Whether the audited operator's purity is what makes the magnitude transferable on that arm is a question one operator cannot settle. As a complement either form leads, by point estimate on both arms, every feature we measured. The stacking pattern replicates on the raw arm, where the erosion add-on's edge over the divergence add-on is itself significant, while on the audited arm that edge is directional. Scored on its own pair, erosion reads 0.87, above the same-pair reference, but this is a near-direct readout of compatibility with the forced answer span rather than a prediction. The committed token's forced-pass probability falls below one half inside the answer window on 99 percent of flipped pairs, so same-pair erosion is an event decomposition, and we class it with the instrumented scores.

J Basis robustness

Divergence is computed in the raw residual-stream basis, and we verified that this choice does not change the conclusions. Residual streams are strongly anisotropic, so we repeated the flip-prediction analysis after an all-but-the-top correction (Mu & Viswanath, 2018), subtracting a mean and removing the top one, five, or ten principal components fitted on states from disjoint prompts. The correction shifts AUROC by -0.002 to -0.008 across depths, always slightly downward and monotone in the number of components removed, while raw displacement magnitudes rise by 13 to 31 percent with one component removed. The dominant common-mode directions therefore carry signal rather than confound it, and the paired contrast already acts as an implicit anisotropy correction, since structure common to both trajectories cancels in the comparison.

K Dose response under compounded lexical perturbation

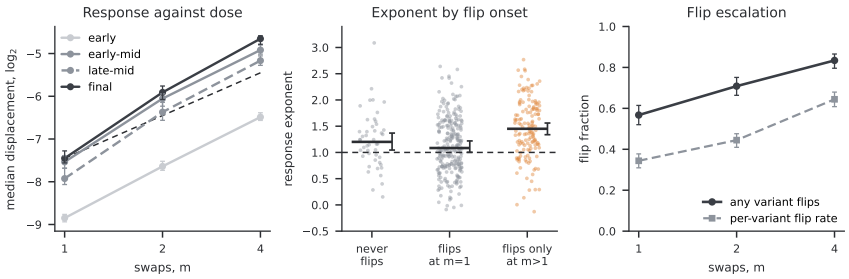

Figure 9: Dose response on Llama-3.1-8B (446 prompts present at all three doses, raw substitution operator at every dose). Left: median teacher-forced displacement against swap count at each depth, on logarithmic axes against a proportional reference (dashed). Center: per-prompt response exponents grouped by flip onset. Right: flip escalation with dose. Single-swap levels sit above the audited arm of Figure 1 because the ungated operator admits harsher swaps. Bars are 95 percent cluster-bootstrap intervals over prompts on each median, each group median, and both proportions.

On Llama-3.1-8B we widen the disturbance to two and four substitutions and ask how displacement scales with dose. A per-swap audit does not establish that a composed perturbation preserves meaning, so these arms use the raw substitution operator at every dose, including the single-swap reference they are paired against, and the axis measures graded lexical disturbance rather than audit-accepted variation. The analysis pairs the 446 prompts present at all three doses.

Displacement grows superlinearly with dose at every depth (Figure 9, left). Doubling from one swap to two multiplies the median displacement by 2.3 to 2.9 depending on depth, and doubling again multiplies it by a further 2.2 to 2.4, both above the factor of two a proportional response

would produce. For cosine divergence that reference is the null of independent displacements adding in quadrature, and coherent displacements would give a factor of four, so the observed factors sit between the two nulls. The raw operator also confounds the axis. Forty-four percent of its single swaps change meaning, so the chance that a variant contains at least one such swap rises from 0.44 at one substitution to about 0.69 at two and 0.90 at four, and meaning-changing swaps carry about three times the divergence of accepted ones (Section 4.3). A mixture of that kind produces superlinear medians on its own, which is why the dose result stays in this appendix. Each prompt's response exponent, the slope of log displacement against log dose at the final depth, has a median of 1.23 with an interquartile range of 0.84 to 1.62, so the population curve averages over heterogeneous individual responses.

The exponent separates prompts by their relationship to the flip boundary. Prompts that resist a single swap but flip at two or four (147 of 446) have a median exponent of 1.45, against 1.08 for prompts that already flip at one swap (two-sided Mann-Whitney $p < 10^{-7}$) and 1.20 for prompts that never flip at any measured dose ($p = 0.009$). Displacement growth accelerates in exactly the group that crosses its boundary within the measured range, which is consistent with a flip marking escape from a locally stable region of state space. We offer that reading as a hypothesis, since swap count is not a calibrated magnitude scale, the grouping conditions on the flip outcome, and the human validation of Section 4.3 shows that compounded raw swaps are increasingly likely to alter the question itself. Two companion observations are consistent with it. The fraction of contracting pairs rises with dose, from 0.92 at one swap to 0.94 at four at the final depth, and flip coverage escalates from 57 percent of prompts with a flipping variant at one swap to 83 percent at four, where the per-variant flip rate reaches 0.64, leaving only 10 percent of prompts that flip at no measured dose.

L CONTROLS FOR ALTERNATIVE EXPLANATIONS

Each main-text result admits an alternative account. This appendix states each account and the control that tests it. Four of the six controls need no new capture; the shared-suffix null and the output-distribution comparison are dedicated captures on Llama-3.1-8B. All use the primary score, the mean teacher-forced cosine divergence of the last-position state at the final depth, the audited captures of every model, the variant-level flip label, and 95 percent cluster-bootstrap intervals over prompts (1,000 resamples). The generating scripts and their outputs are released with the capture.

Length, not meaning. A one-word substitution can change the prompt's subword count, which alters the tokenized context as well as its lexical content. Table 9 gives the prompt-length delta computed with each model's own tokenizer through the capture's chat template. The delta is zero for 79 to 90 percent of pairs and never exceeds two tokens in magnitude. Flip AUROC on length-preserving pairs is statistically indistinguishable from AUROC on length-changing pairs in every model, so length change does not carry the signal, although absolute length change correlates weakly with the score and flip rates run mildly higher in the length-changing strata.

model	share $\Delta = 0$	AUROC, $\Delta = 0$	AUROC, $\Delta \neq 0$	difference	$\rho(\lvert\Delta\rvert, \text{score})$
Llama-3.1-8B	0.81	0.792 [0.751, 0.827]	0.798 [0.723, 0.860]	-0.006 [-0.082, +0.070]	+0.103 [+0.048, +0.156]
Gemma-3-4B	0.90	0.764 [0.731, 0.795]	0.754 [0.663, 0.836]	+0.009 [-0.078, +0.109]	+0.053 [-0.004, +0.104]
Qwen3-8B	0.81	0.707 [0.665, 0.747]	0.725 [0.652, 0.792]	-0.018 [-0.095, +0.063]	+0.147 [+0.094, +0.194]
Gemma-3-27B	0.90	0.783 [0.739, 0.827]	0.780 [0.657, 0.883]	+0.003 [-0.118, +0.129]	+0.038 [-0.019, +0.087]
Qwen3-32B	0.81	0.766 [0.729, 0.803]	0.818 [0.762, 0.873]	-0.052 [-0.117, +0.013]	+0.126 [+0.069, +0.176]
Llama-3.3-70B	0.79	0.773 [0.662, 0.869]	0.736 [0.573, 0.877]	+0.037 [-0.122, +0.208]	+0.142 [+0.070, +0.212]

Table 9: Tokenization-length control. Δ is the perturbed minus clean prompt token count. AUROC is for the flip target within each stratum, the difference is the $\Delta = 0$ stratum minus the $\Delta \neq 0$ stratum, and ρ is the Spearman correlation of $|\Delta|$ with the primary score across pairs.

Imbalance flatters AUROC. AUROC is forgiving under class imbalance, and Llama-3.3-70B flips on only 8 percent of pairs. Table 10 reports average precision (AP) of the primary score against the best confidence baseline from the same capture, and splits flips by the correctness of the clean and perturbed answers. AP runs 0.40 to 0.61 against base rates of 0.08 to 0.34, a lift of 1.8 to 4.8 over chance, and the AP gain over the baseline is credible on three models. Flagging the top 20 percent of pairs by score yields precision 0.26 to 0.69 and recall 0.39 to 0.62, each with its interval in Table 10.

On four of six models most flips are changes between two wrong answers (29 to 70 percent of flips across the grid) rather than correctness losses, and no flip is a pure normalization difference, since the flip label already negates the normalized-match rule used for grading. The score is at its best on the class that matters, separating correct-to-wrong flips from non-flips at AUROC 0.76 to 0.87 and beating the confidence baseline on all six models.

				top 20 percent flagged		share of flips		correct-to-wrong vs. non-flip	
model	flip rate	AP, divergence	AP gain vs. baseline	precision	recall	c→w	w→w′	AUROC, divergence	AUROC, baseline
Llama-3.1-8B	0.225	0.571 [0.508, 0.637]	+0.122 [+0.053, +0.188]	0.60 [0.51, 0.68]	0.53 [0.48, 0.58]	0.27	0.50	0.831 [0.776, 0.883]	0.720
Gemma-3-4B	0.344	0.614 [0.560, 0.675]	+0.127 [+0.066, +0.190]	0.69 [0.62, 0.75]	0.40 [0.36, 0.44]	0.16	0.70	0.779 [0.721, 0.830]	0.620
Qwen3-8B	0.260	0.505 [0.447, 0.567]	+0.054 [-0.017, +0.116]	0.50 [0.44, 0.58]	0.39 [0.34, 0.43]	0.25	0.61	0.757 [0.697, 0.819]	0.604
Gemma-3-27B	0.131	0.412 [0.337, 0.491]	+0.176 [+0.099, +0.252]	0.36 [0.30, 0.42]	0.56 [0.48, 0.63]	0.38	0.46	0.866 [0.807, 0.922]	0.660
Qwen3-32B	0.222	0.483 [0.419, 0.558]	-0.016 [-0.093, +0.073]	0.53 [0.45, 0.60]	0.47 [0.43, 0.53]	0.27	0.57	0.801 [0.750, 0.853]	0.702
Llama-3.3-70B	0.083	0.395 [0.271, 0.531]	+0.100 [-0.050, +0.231]	0.26 [0.18, 0.34]	0.62 [0.48, 0.76]	0.46	0.29	0.782 [0.632, 0.913]	0.680

Table 10: Average precision and flip taxonomy for the flip target. Precision and recall are for the top 20 percent of pairs by primary score, with 95 percent cluster-bootstrap intervals over prompts. c→w is the share of flips in which a correct clean answer becomes wrong, and w→w′ the share in which a wrong answer becomes a different wrong answer. The last two columns score correct-to-wrong flips against non-flips with the other flip classes removed.

The secondary bump is the output format. The median teacher-forced divergence curve shows a bump at steps 9 to 13 in every model. Our format asks for a short answer followed by a justification, so we located, for every clean generation, the step at which the decoded prefix first passes the justification marker, and compared it with the per-prompt divergence peak searched over steps 5 to 20 (Table 11). The boundary sits at median step 5 to 7 and falls inside the bump window for 76 to 95 percent of prompts, and the per-prompt peak trails the boundary by a median of 5 to 6 steps. Realigning every curve on its own boundary more than doubles the peak-over-trough prominence and exposes a near-zero trough one step before the boundary, where the template token is nearly deterministic. The bump is therefore the template handoff into the justification, a format feature rather than a model-generic dynamic, and the main text no longer leans on it.

						peak-over-trough	
model	boundary step (IQR)	bump step	boundary in window	peak minus boundary	peak 0 to 6 after	unaligned	realigned
Llama-3.1-8B	6 (5 to 8)	11	0.87	+5 [+4, +5]	0.48 [0.44, 0.52]	2.33	6.92
Gemma-3-4B	6 (5 to 7)	11	0.93	+6 [+5, +7]	0.47 [0.43, 0.51]	8.06	22.46
Qwen3-8B	6 (5 to 7)	11	0.86	+6 [+6, +7]	0.46 [0.42, 0.50]	3.09	6.34
Gemma-3-27B	5 (5 to 7)	9	0.76	+5 [+4, +5]	0.57 [0.53, 0.62]	1.62	7.47
Qwen3-32B	7 (6 to 8)	13	0.95	+6 [+6, +7]	0.44 [0.39, 0.48]	3.09	6.03
Llama-3.3-70B	5 (5 to 6)	11	0.76	+6 [+5, +7]	0.47 [0.41, 0.54]	4.54	10.09

Table 11: Alignment of the secondary bump with the answer-to-justification boundary. Boundary in window is the share of prompts whose boundary falls in steps 5 to 20. Prominence is the ratio of the median-curve peak to its preceding trough, unaligned in absolute steps and realigned to each prompt's boundary.

Teacher forcing absorbs perturbations by construction. Under condition (b) the clean and perturbed computations consume an ever-longer identical continuation, so the one differing prompt word sits ever further back in a context that is otherwise the same, and the two last-position states would converge whether or not the swap changed anything. On that account the 92 to 97 percent contraction of Section 5 could be dilution of the differing prefix, not a property of the model's response. To measure how much of it is, we captured, on Llama-3.1-8B, five semantically inert perturbations of every prompt through the same forced pass against the same clean references: three one-word swaps in the instruction scaffold, far from the question, and two changes at the position of audited variant one that alter tokenization without changing a word, a case flip and a double space (2,462 pairs, no failures). Table 12 compares the two arms conditional on initial displacement d_0, the mean divergence over the first three steps, using the residual fraction (mean divergence over the last ten of sixty steps divided by d_0) and the share of pairs whose whole-trajectory contraction slope is negative, the definition of Section 5. Inert perturbations contract in 95 to 98 percent of pairs against 90 to 97 percent for audited swaps across the four depths, so contraction is mostly a property of the shared continuation, and the arms differ more in how steeply they decay than in whether they decay. Audited swaps differ in two ways. Their initial displacement is two to eight times larger at every

depth, and among pairs of small initial displacement they retain a residual fraction two to four times that of inert perturbations at every depth (differences 0.20 to 0.47, intervals excluding zero). Among pairs of large initial displacement the excess shrinks with depth and reverses at the final depth, where large disturbances relax to a similar or smaller fraction regardless of kind. At the final depth and small initial size, non-flipped audited pairs retain 0.42 (interval 0.36 to 0.47) against 0.22 (0.22 to 0.23) for inert pairs, so even the audited swaps that do not flip persist above the null. The flipped pairs of that cell retain 0.62 on 16 pairs over 14 prompts, an interval of 0.29 to 0.94 that supports no ordering against the non-flipped pairs. Figure 10 shows both views. Pooled over all pairs the two decay curves are close, and the audited curve ends lower because it pools larger disturbances, which relax to smaller fractions; the matched comparison on the right is where the arms separate. Contraction is therefore mostly a control property of the forced pass, and what the assay measures above it is the additional size and persistence associated with the audited lexical substitution.

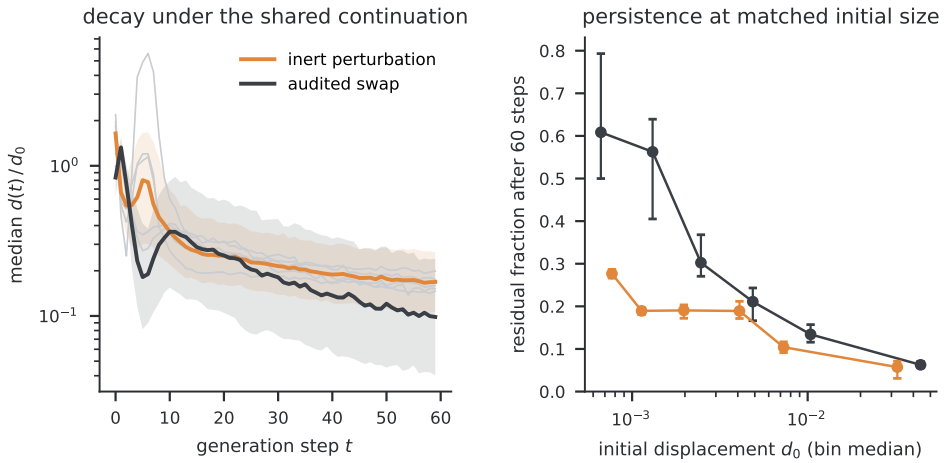

Figure 10: The shared-suffix null on Llama-3.1-8B, final depth. Left: median teacher-forced divergence over generation steps, each pair normalized by its own initial displacement, for audited swaps (dark) and inert perturbations (orange, thin lines for the five null families). Right: median residual fraction after sixty steps against initial displacement, by pooled-quantile bin, with prompt-level cluster-bootstrap intervals. The shaded bands on the two pooled curves are interquartile ranges across pairs, and the thin gray lines are the five inert families rather than an interval. At matched small initial displacement, audited swaps retain two to four times the residual of inert perturbations.

	median d_0		residual fraction, small d_0			residual fraction, large d_0			contracting	
depth	swap	null	swap	null	difference	swap	null	difference	swap	null
early	0.0026	0.0011	0.52	0.14	+0.371 [+0.282, +0.461]	0.17	0.15	+0.025 [+0.008, +0.043]	0.97 [0.96, 0.98]	0.98 [0.97, 0.99]
early-mid	0.0059	0.0015	0.64	0.17	+0.472 [+0.369, +0.556]	0.24	0.13	+0.108 [+0.078, +0.132]	0.91 [0.88, 0.93]	0.96 [0.94, 0.97]
late-mid	0.0059	0.0010	0.59	0.16	+0.429 [+0.353, +0.509]	0.14	0.13	+0.008 [-0.012, +0.035]	0.90 [0.87, 0.92]	0.95 [0.94, 0.96]
final	0.0121	0.0015	0.42	0.22	+0.201 [+0.138, +0.248]	0.10	0.15	−0.042 [-0.060, -0.028]	0.94 [0.92, 0.96]	0.97 [0.96, 0.98]

Table 12: Shared-suffix null on Llama-3.1-8B. Median initial displacement d_0 per arm, the residual fraction after sixty steps for audited swaps and inert perturbations pooled over the three smallest and the three largest pooled-quantile bins of d_0, with a prompt-level cluster-bootstrap interval on the swap-minus-null difference, and the share of contracting pairs with its interval. The ratio of median initial displacement, audited over inert, runs from 2.3 (interval 2.1 to 2.5) at the earliest depth to 7.8 (6.7 to 8.7) at the final depth. Pooled over all pairs rather than within bins, the flipped-minus-inert difference in residual fraction changes sign with depth, from $+0.08$ ($+0.04$ to $+0.12$) at the earliest depth to -0.12 (-0.14 to -0.10) at the final depth, because the audited arm pools larger initial displacements, so we base the interpretation on the matched comparison within bins. Last-token state, condition (b).

The trajectory adds little over its first position. Table 14 gives the flip AUROC of the trajectory mean against the first and the last teacher-forced position, with the paired bootstrap difference. The

first position alone reaches 0.69 to 0.77, and the trajectory mean adds a median 0.03 with an interval excluding zero on three of six models, while against the last position the median gain is 0.08 and five of six intervals exclude zero. Table 13 makes the same point from the other side. Excluding the answer window from the score costs 0.017 to 0.033 AUROC on the six models, every interval clear of zero, and leaves the paired difference between the mean and the first position spanning zero on five of six models (Gemma-3-4B is the exception at $+0.035$, interval 0.008 to 0.063), so the trajectory's gain over its first position was the window, where the score has direct access to the forced clean answer span. The forced justification restates the clean answer, so positions after the window can still register disagreement on flipping pairs. The first-position score sees only the shared format token that opens every continuation and no answer content, and it matches the window-excluded mean on five of six models, which bounds that residual. The paired absolute log-odds erosion of the forced clean tokens over that window reads 0.89 to 0.97, a near-direct readout of compatibility with the answer span, and over the positions after it 0.66 to 0.74, which the window-excluded latent score leads on every model by point estimate and on three with intervals excluding zero. Much of the flip-predictive signal is already present at the prompt boundary, and the full curve's margin over its first position is not evidence that later positions predict.

Model	Latent, full	Latent, first position	Latent, window excluded	Paired output, window excluded	Excluding the window	Latent minus paired output
Llama-3.1-8B	0.792	0.752	0.758 [0.719, 0.797]	0.724 [0.683, 0.761]	-0.033 [-0.048, -0.022]	+0.034 [+0.010, +0.060]
Gemma-3-4B	0.762	0.707	0.742 [0.710, 0.773]	0.706 [0.673, 0.737]	-0.020 [-0.026, -0.015]	+0.036 [+0.019, +0.054]
Qwen3-8B	0.713	0.691	0.682 [0.644, 0.716]	0.663 [0.626, 0.699]	-0.031 [-0.043, -0.020]	+0.020 [-0.008, +0.046]
Gemma-3-27B	0.782	0.772	0.757 [0.716, 0.802]	0.744 [0.703, 0.786]	-0.025 [-0.032, -0.019]	+0.013 [-0.010, +0.034]
Qwen3-32B	0.778	0.753	0.753 [0.721, 0.787]	0.741 [0.705, 0.778]	-0.025 [-0.033, -0.017]	+0.012 [-0.018, +0.041]
Llama-3.3-70B	0.769	0.711	0.752 [0.659, 0.835]	0.701 [0.624, 0.781]	-0.017 [-0.031, -0.007]	+0.051 [+0.008, +0.102]

Table 13: The answer window and the paired output comparator. Flip AUROC of the primary latent score over the full continuation, at the first position, and with the eight-position answer window excluded, beside the paired absolute log-odds erosion of the forced clean tokens with the same window excluded, then the paired change from excluding the window and the paired margin of the window-excluded latent score over the window-excluded paired output score. Intervals are 95 percent prompt-cluster bootstraps.

	flip AUROC			trajectory minus position	
model	trajectory	first position	last position	vs. first	vs. last
Llama-3.1-8B	0.792	0.752	0.733	+0.040 [+0.011, +0.066]	+0.058 [+0.028, +0.087]
Gemma-3-4B	0.762	0.707	0.684	+0.056 [+0.030, +0.083]	+0.078 [+0.051, +0.108]
Qwen3-8B	0.713	0.691	0.622	+0.022 [-0.006, +0.051]	+0.091 [+0.055, +0.130]
Gemma-3-27B	0.782	0.772	0.703	+0.010 [-0.016, +0.036]	+0.080 [+0.046, +0.115]
Qwen3-32B	0.778	0.753	0.682	+0.024 [-0.005, +0.055]	+0.096 [+0.064, +0.128]
Llama-3.3-70B	0.769	0.711	0.747	+0.058 [+0.010, +0.108]	+0.021 [-0.028, +0.067]

Table 14: Flip AUROC of the trajectory mean against single positions of the same teacher-forced curve, with the paired cluster-bootstrap difference. Positive values favor the trajectory.

Paired output distributions match the latent score only inside the answer window. The output head sees the same forced path, so the closest output-space comparator to a hidden-state divergence is a divergence of the two full next-token distributions at every position, computed from the same two forward passes with no hidden-state access. On Llama-3.1-8B we compared the clean and perturbed next-token distributions under the clean continuation at every generated position by Jensen-Shannon divergence, symmetric KL, total variation, and top-1 disagreement (Table 15). Averaged over the trajectory, the output divergence and the latent score are at parity (0.80 against 0.79, a paired difference whose interval spans zero, Spearman correlation 0.84 across pairs), and stacking the two adds nothing. The parity comes from the answer window. Excluding the first eight positions from both scores, the latent score reads 0.76 against 0.72 for the output divergence (paired difference 0.034, interval 0.010 to 0.060, Table 13). Restricted to those eight positions, the output divergence reaches 0.88, the same value as the signed erosion score and for the same reason. In that window the score is a near-direct readout of compatibility with the forced answer span rather than a prediction, and we class it with the instrumented scores. At the first position the ordering reverses. The output distribution there is pinned by the answer-first format, so its divergence discriminates flips at only 0.64, while the first-position latent state reaches 0.75 (the previous control). The output head substitutes for the state

once the forced path carries the answer. It does not at the prompt boundary or across depth, which is where the instrument's claims are made.

score	flip AUROC	difference from latent
latent divergence, trajectory mean (primary)	0.792 [0.756, 0.825]	
output JS divergence, first position	0.640 [0.597, 0.684]	−0.152 [-0.193, -0.112]
output JS divergence, trajectory mean	0.799 [0.766, 0.831]	+0.007 [-0.014, +0.027]
output JS divergence, maximum	0.780 [0.748, 0.813]	−0.012 [-0.043, +0.019]
symmetric KL, trajectory mean	0.794 [0.763, 0.827]	+0.003 [-0.019, +0.026]
total variation, trajectory mean	0.785 [0.749, 0.817]	−0.007 [-0.029, +0.013]
top-1 disagreement rate	0.770 [0.733, 0.804]	−0.022 [-0.053, +0.011]
output JS divergence, first eight positions	0.876 [0.851, 0.903]	+0.085 [+0.056, +0.115]
signed log-odds erosion, first eight positions	0.873 [0.841, 0.902]	+0.081 [+0.039, +0.123]

Table 15: Flip AUROC on Llama-3.1-8B of paired next-token distribution divergences along the forced continuation, against the primary latent score, with 95 percent cluster-bootstrap intervals over prompts and the paired difference from the latent score. Forced-token log-probabilities from this capture reproduce the stored clean values to a median absolute difference of 0.001.

M LIMITATIONS IN FULL

Decoding is greedy outside the resampling control, so we measure stability of the deterministic path, not of sampled decoding. The primary operator is a single counter-fitted synonym substitution, audited but not general. The language-model audit passes 14 percent of swaps that a human rejects, the human sample covers 360 of 1,376 variants with a second annotator on 120, and the headline analyses on the human-confirmed subset (Section 4.3, Appendix D) carry intervals about two to three times wider than on the full set. The screening comparison of Section 6.1 was run on the audit-accepted population and not repeated under human filtering, so the ranking between screens is established for that population only. The showcased swaps in Appendix C are drawn from the strict subset of the human audit. The contraction result is read against an inert-perturbation null on Llama-3.1-8B only; the null shows that contraction is mostly the shared continuation at work, and the residual above the null is reported as an impulse response, not as evidence of an attractor. Substitutions change the prompt token count in 10 to 21 percent of pairs, and although flip AUROC does not differ between length-preserving and length-changing pairs, absolute length change correlates weakly with the divergence score (Spearman 0.04 to 0.15), a residual channel we disclose rather than remove. The secondary divergence bump is tied to the answer-to-justification handoff of our output format and should not be read as a model-generic dynamic. Beyond these, the dose axis reaches four substitutions without the per-swap audit, paraphrase-level perturbation remains future work, and every perturbation is lexical, English, and applied to short-form factual questions. Flip and correctness labels come from normalized string matching, and a spot audit found equivalence misses at roughly the one percent level, which mislabel stable answers as flips. We have not tested whether those misses are independent of the score, so we do not claim that they only attenuate the AUROCs. The flip target reads only the short-answer span, and onset inherits a circularity from the answer-first format, so we exclude it from the headline claims. The population is six dense instruction-tuned transformers between 4B and 70B parameters, with mixture-of-experts architectures excluded because expert routing changes the semantics of hidden-state comparison. We chose the absolute-erosion form after seeing the Llama-3.1-8B results, so its out-of-sample replication on the other five models inherits that ordering, and the stacked screen shows no gain on Llama-3.3-70B, whose 245-prompt, eight-percent-flip cell supplies the least evidence here. Teacher forcing runs in one direction, the clean continuation through the perturbed prompt, and we have not run the symmetric check that forces the perturbed continuation through the clean prompt. The study is observational throughout, and we make no causal or mechanistic claims about why particular prompts sit in the fragile tail.