SAGE: Safety-Boundary Atlas for Auditing Language Models Beyond Endpoint Labels

Anonymous ACL submission

Abstract

Safety evaluations of language models are commonly summarized by endpoint labels such as refusal, compliance, harmfulness, or jailbreak success. Although useful for deployment-facing reporting, these labels conflate several distinct effects: whether an intervention moved an internal safety boundary, redirected the final answer, degraded generation quality, or damaged benign behavior. We formalize this failure mode as the *endpoint-collapse problem* and introduce SAGE (Safety Atlas for Geometric Evaluation), a white-box diagnostic framework for auditing language-model safety boundaries beyond endpoint labels. SAGE builds prompt-wise, native-calibrated functional atlases over layer depth and generation phase using operational residual-state readouts for harmfulness, refusal/safety, procedural compliance, and benign continuation. It then fits layer–phase functionals to select diagnostic source-window probes and reports intervention effects along four separated axes: internal functional movement, endpoint behavior, output quality, and preservation cost.

Across 1,368 native calibration prompts and 555 held-out JailbreakBench prompts on Qwen2.5 scales, SAGE shows that formula-selected source windows induce substantially larger model-local harmfulness/refusal functional movement than matched random controls. At the same time, endpoint observability is strongly benchmark-dependent: JailbreakBench exposes boundary cases for posterior audit, whereas StrongREJECT is saturated for this model family. SAGE is not a deployable defense, a jailbreak method, or a claim of discovered circuitry; it is a measurement protocol for determining whether safety interventions move the intended internal geometry rather than merely changing surface labels.

1 Introduction

Modern language-model safety evaluations often reduce behavior to endpoint labels: a response is marked as refusal, compliance, harmfulness, or jailbreak success. These labels are indispensable for deployment-facing evaluation, but they are also severely lossy. Two interventions with identical endpoint outcomes can have moved very different internal states; conversely, an intervention can substantially alter internal safety geometry while remaining invisible at the final-answer level. A safer-looking endpoint can also be purchased by incoherent generation, degraded instruction following, or damage to benign behavior. Endpoint-only evaluation therefore makes it difficult to determine whether an intervention actually affects the intended safety boundary.

We call this the *endpoint-collapse problem*: final-answer labels conflate internal functional movement, endpoint redirection, output quality, and preservation cost. This problem is especially acute for activation-space intervention studies. A steering result may change a label without moving the intended source geometry, or it may move a meaningful internal boundary while endpoint behavior remains gated by downstream decoding, benchmark saturation, or model-family-specific refusal policies. Treating all of these cases as a single success rate hides the mechanism that the intervention was meant to test.

This paper argues that safety-boundary interventions should be reported as a diagnostic audit rather than as a single endpoint success rate. SAGE does not aim to provide a deployable safety intervention; it provides a reporting framework for separating four quantities: internal functional movement, endpoint redirection, output quality, and preservation cost. This separation turns endpoint labels from the sole target of evaluation into one axis of a broader safety-boundary audit.

Prior work gives strong evidence that activation directions can influence behavior. Representation engineering, activation engineering, refusal-vector methods, and activation patching all suggest that

internal states contain steerable safety information (Zou et al., 2023; Turner et al., 2023; Arditi et al., 2024; Zhang and Nanda, 2023). However, a steering result by itself does not tell us whether the intended source window moved, whether an endpoint changed for the right reason, or whether useful behavior was preserved. SAGE is designed as a measurement layer around such interventions, not as a competing defense.

SAGE maps each prompt into a native-calibrated functional atlas over layer depth and generation phase. First, it builds operational residual-state readouts for harmfulness, refusal/safety, procedural compliance, and benign continuation. These readouts define model-local coordinates rather than universal semantic scales. Second, SAGE fits trajectory summaries over these coordinates, including harmfulness-minus-refusal margins, risk summaries, ridges, persistence terms, and source windows. Third, it uses the fitted formulas to choose layer–phase source cells and intervention directions. Each probe is then audited along the four axes above instead of treating endpoint conversion as the only outcome.

This structure leads to three central empirical observations. First, endpoint-observable benchmark regimes matter: held-out JailbreakBench prompts produce usable Qwen2.5 boundary cases, while StrongREJECT is saturated in this family and therefore less informative for endpoint-transition analysis. Second, formula-selected source windows can move fitted harmfulness/refusal functionals much more strongly than matched random controls across Qwen2.5 scales. Third, these functional movements and endpoint outcomes must be interpreted separately: SAGE is a diagnostic atlas for measuring internal safety geometry, not a claim of endpoint-control superiority.

Our contributions are:

- **Endpoint-collapse formulation.** We identify the endpoint-collapse problem in safety intervention evaluation: final-answer labels conflate internal functional movement, endpoint redirection, output quality, and preservation cost.

- **Safety-boundary atlas construction.** We introduce SAGE (Safety Atlas for Geometric Evaluation), a white-box diagnostic framework that constructs native-calibrated harmfulness, refusal/safety, procedural-compliance, and benign-continuation fields over layer depth and generation phase.

- **Four-axis source-window audit protocol.** We propose a formula-driven protocol for selecting diagnostic source-window probes and auditing their effects along four separated axes, preventing endpoint behavior from being mistaken for internal safety-boundary movement.

- **Cross-scale empirical evidence.** Across Qwen2.5 scales, SAGE shows that formula-selected source windows induce substantially larger harmfulness/refusal functional movement than matched random controls, and that endpoint observability differs sharply across benchmarks.

Figure 1 summarizes this framing: endpoint labels are compressed observations of a richer prompt-wise functional geometry, and intervention results are audited by whether internal movement, endpoint behavior, output quality, and preservation cost separate.

2 Related Work

Safety benchmarks and endpoint labels. JailbreakBench and StrongREJECT provide standardized ways to evaluate adversarial or harmful behavior (Chao et al., 2024; Souly et al., 2024). These benchmarks are indispensable for measuring final-answer safety, but endpoint labels alone do not expose the internal trajectory that produced an answer. SAGE uses benchmarks differently: it asks whether a benchmark supplies an endpoint-observable boundary regime for intervention diagnosis. A saturated benchmark can still test robustness, but it becomes a weak source of endpoint-transition evidence.

Activation steering and representation editing. Representation engineering and activation engineering show that activation-space directions can steer model behavior (Zou et al., 2023; Turner et al., 2023). Refusal-vector work further suggests that refusal can be mediated by a low-dimensional direction in some models (Arditi et al., 2024). Much of this literature reports behavioral change or endpoint shifts, while fewer evaluations require all four pieces of evidence at once: source-local functional movement, endpoint behavior, output quality, and preservation cost. SAGE does not claim to outperform these steering methods. It supplies a diagnostic audit protocol around them: did the intended functional geometry move, did endpoint behavior follow, did output quality survive, and what

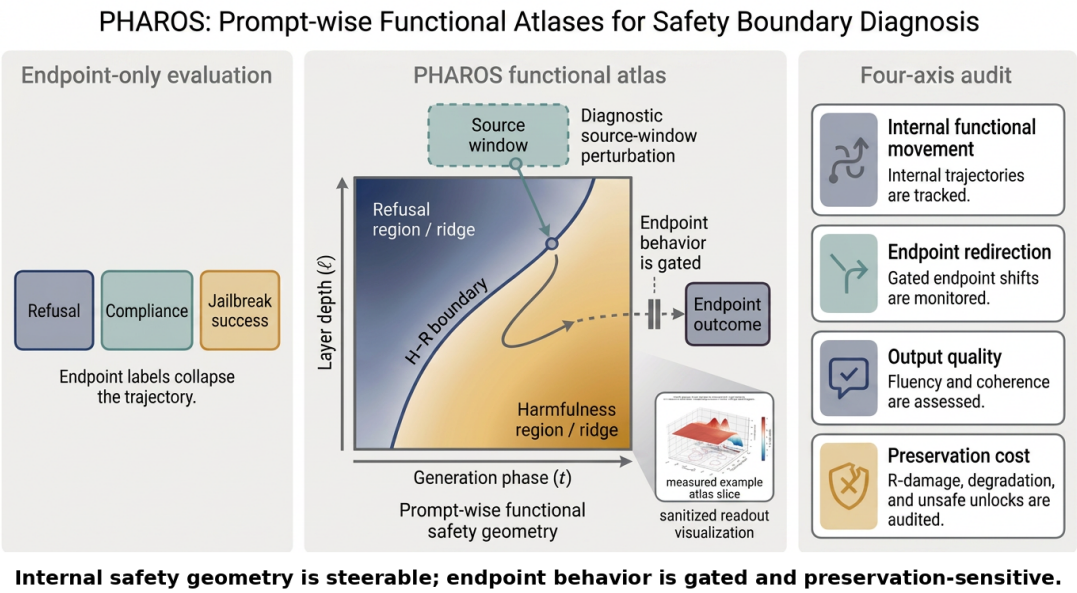

Figure 1: Conceptual framing of SAGE. Endpoint-only evaluation collapses model safety behavior into final labels, whereas SAGE represents prompt-wise functional safety geometry over layer depth and generation phase. The central panel uses a measured H–R atlas slice from fitted readouts; source-window boxes and arrows are explanatory overlays. Diagnostic source-window perturbations are audited along four separated axes–internal functional movement, endpoint redirection, output quality, and preservation cost–highlighting that endpoint behavior can remain gated even when internal geometry moves.

preservation cost appeared? This makes SAGE a natural wrapper for future side-by-side evaluations of refusal-direction, activation-engineering, SAE-feature, or state-dependent steering baselines under the same reporting axes.

Mechanistic interpretability and causal localization. The readout construction is related to logit-lens and tuned-lens analyses (Belrose et al., 2023). Sparse autoencoder and monosemantic-feature work could refine the coarse functional fields used here (Cunningham et al., 2023; Templeton et al., 2024). Activation patching and model editing motivate the source-window view, but also caution against overclaiming causal localization (Meng et al., 2022; Zhang and Nanda, 2023). SAGE therefore treats layer–phase cells as operational source windows and fitted energy/risk summaries, not as proven circuit components or continuous-time stability certificates. SAE-based methods may provide deeper feature localization; SAGE instead focuses on an audit protocol for testing source-level functional movement before interpreting endpoint behavior.

Safety auditing and evaluation protocols. SAGE belongs most naturally to safety auditing and evaluation methodology. Its central goal is not to discover a universal safety direction, maximize jailbreak success, or certify a deployable defense. Instead, it decomposes what endpoint metrics conflate: whether the intended internal geometry moved, whether endpoint behavior changed, whether quality remained acceptable, and whether benign behavior was preserved.

3 Method

As shown in Figure 2, SAGE consists of two stages: functional atlas construction followed by formula-driven intervention audit. Stage I is a modeling stage: construct model-local functional readouts, calibrate them on native prompts, and fit scalar energy/risk formulas over prompt trajectories. Stage II is an intervention stage: use those fitted formulas to rank source cells, perturb the selected cells, and audit the result along separated outcome axes.

SAGE is designed around three principles. First, *model-local calibration*: safety coordinates should be comparable within a model without claiming universal semantic scales across models. Second, *prompt-wise geometry*: safety-boundary dynamics should be localized over layer depth and generation phase rather than compressed into a single endpoint label. Third, *audit separation*: internal movement, endpoint behavior, quality, and preservation should be reported separately.

Stage I: functional readouts. For each field $a \in \{\mathrm{H}, \mathrm{R}, \mathrm{C}, \mathrm{B}\}$, SAGE defines a concept-token

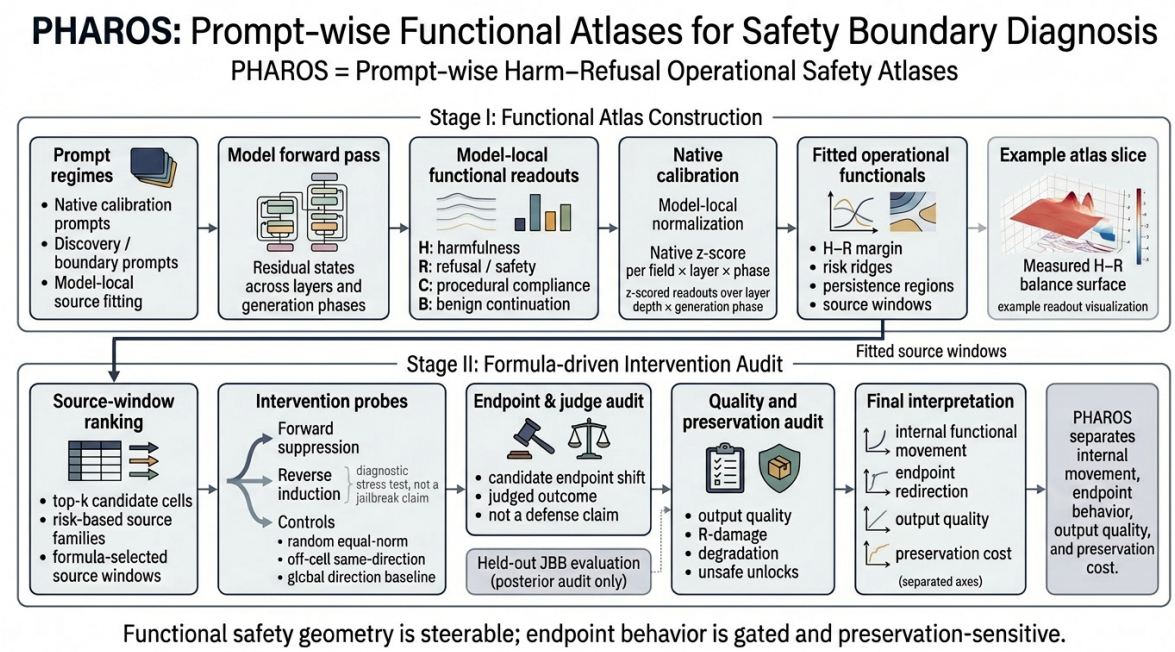

Figure 2: Overview of the SAGE pipeline. Stage I constructs prompt-wise functional atlases from model-local readouts over layer depth and generation phase, using native calibration to fit operational harmfulness/refusal/procedural-compliance/benign-continuation functionals and source windows. Stage II converts fitted source windows into intervention probes and audits outcomes along four separated axes: internal functional movement, endpoint behavior, output quality, and preservation cost. Held-out JailbreakBench prompts are used only for posterior endpoint/judge audit, not for source fitting or intervention selection.

set T_a and a neutral background set T_0. The four fields correspond to harmfulness, refusal/safety, procedural compliance, and benign continuation. Let $h_{\ell,t}(p)$ be the residual state for prompt p at layer ℓ and generation-phase slot t, and let e_u be the model-local output-readout row for vocabulary token u. We compute a vocabulary-logit readout,

$$z_u(\ell, t; p) = \langle h_{\ell,t}(p), e_u \rangle + b_u, \tag{1}$$

where b_u is omitted for architectures without output bias. The raw field score is a background-centered concept contrast:

$$\begin{aligned} r_a(\ell, t; p) = &\operatorname*{logsumexp}_{u \in T_a} z_u(\ell, t; p) \\ &- \operatorname*{logsumexp}_{u \in T_0} z_u(\ell, t; p). \end{aligned} \tag{2}$$

Native prompts define model-local means and variances, yielding calibrated scores

$$\tilde{r}_a(\ell, t; p) = \frac{r_a(\ell, t; p) - \mu^{native}_{a,\ell,t}}{\sigma^{native}_{a,\ell,t} + \epsilon}. \tag{3}$$

This readout is not a trained endpoint classifier. It is an operational coordinate system for comparing prompt trajectories within the same model. The harmfulness and refusal/safety fields are the primary boundary axes; procedural compliance and benign continuation are auxiliary axes used to separate rule-following and ordinary continuation behavior from safety-specific movement, and to make preservation costs visible rather than silently absorbed into endpoint labels. Additional readout construction details, token filtering rules, native-calibration protocol, and functional-scope clarifications are provided in Appendices A and C.

Stage I: operational energy/risk fitting. From the calibrated surfaces, SAGE fits trajectory-level functionals. A generic harmfulness-minus-refusal energy margin over a source or phase region Ω is

$$F(p) = \mathbb{E}_{c \in \Omega}[\tilde{r}_{\mathrm{H}}(c; p) - \tilde{r}_{\mathrm{R}}(c; p)]. \tag{4}$$

The broader risk functional combines late harmfulness energy, harmfulness-minus-refusal excess, ridge persistence, harmfulness/refusal coupling, velocity-like finite differences, and preservation terms. The sparse weights are model-local and fit on discovery/source-family splits; the resulting scalar is used for source-cell ranking and audit design, not as a universal risk predictor. We use the term Lyapunov-style only as an operational analogy: the fitted scalar summarizes movement toward or away from a risky functional region, but it is not a continuous-time stability theorem, a PDE claim, or a certified Lyapunov function.

Component	Operational quantity	Estimated from	Used for
H/R/C/B readout fields	Native-calibrated surfaces $\tilde{r}_a(\ell, t; p)$ over layer and generation phase	1368 native prompts and model-local concept-token sets	Prompt-wise functional atlas construction
Energy margin	$F(p) = \mathbb{E}_{c\in\Omega}[\tilde{r}_{\mathrm{H}}(c; p) - \tilde{r}_{\mathrm{R}}(c; p)]$	Native, discovery, and wrapped prompt trajectories	Boundary ranking and H/R excess measurement
Risk functional	Sparse weighted score over late H energy, H–R excess, persistence, coupling, and velocity-like changes	Source-family discovery splits	Source-cell ranking before intervention
Source windows	Layer–phase cells with high fitted risk, ridge persistence, or R-preserving H excess	Atlas surfaces and prompt-local rankings	Forward suppression, reverse induction, and off-cell controls
Preservation gate	R-field damage, IO-fitness loss, unsafe unlocks, and degradation counts	Held-out audits and judge summaries	Reject high-cost endpoint shifts

Table 1: SAGE Stage I functional model components and how they feed Stage II interventions. The Lyapunov-style language denotes an operational scalar energy/risk model over residual-state trajectories; it is not a theorem of continuous-time stability.

Stage II: formula-driven intervention probes. Each layer–phase slot $c = (\ell, t)$ is a candidate source cell. SAGE ranks cells using the fitted prompt-local functionals: late harmfulness persistence, harmfulness-minus-refusal excess, high continuous risk, and refusal-preserving harmfulness ridges. For a source family s, the top cells are

$$C_s^k(p) = \mathrm{TopK}_c \; r_s(p, c), \tag{5}$$

and interventions perturb these cells as $h'_c = h_c + \alpha\delta_c^s$ for $c \in C_s^k(p)$. Matched controls use random equal-norm perturbations or off-cell same-direction perturbations. The evaluation reports four axes: internal functional movement, endpoint redirection, input-output quality, and preservation cost. SAGE therefore tests whether a fitted source-window probe behaves differently from norm- and locality-matched controls, without claiming that the source window is a discovered circuit. Full intervention schedules, matched controls, endpoint-audit rules, quality checks, and preservation-cost diagnostics are detailed in Appendices D and E.

Audit axis	What SAGE measures	Why endpoint-only misses it
Functional movement	H/R atlas and fitted-risk shift	Internal geometry may move without label change
Endpoint redirection	Refusal/compliance or harmful-like label	Labels may change for the wrong reason
Quality	IO-fitness, coherence, malformed outputs	Apparent safety can reflect degradation
Preservation	Benign-continuation and unsafe-unlock cost	Safety shifts may damage useful behavior

Table 2: Four-axis SAGE audit decomposition. Endpoint behavior is one reported axis rather than the sole criterion for intervention success.

4 Experiments

Table 3 summarizes the prompt regimes and endpoint-observability split. Native prompts calibrate model-local coordinates; discovery and wrapped prompts expose source morphology; held-out JailbreakBench prompts provide endpoint-observable posterior evaluation. The main analyses use the 1368-prompt native baseline manifest, sanitized endpoint summaries over 555 held-out JailbreakBench prompts, and Qwen2.5 cross-scale functional-movement audits.

Experiments use Qwen2.5 scales as the main model family. The main text focuses on benchmark observability, Qwen2.5 cross-scale source-window functional movement, cached readout robustness, and two compact appendix-promoted audits: a preservation-positive endpoint source result and a non-Qwen functional stress test. To support reproducibility, all sanitized code, configuration files, experiment scripts, aggregate result tables, package manifests, checksum metadata, and experiment details will be released on GitHub upon publication. The release excludes raw harmful prompts, raw model generations, private judge logs, unsafe generated text, credentials, and endpoint access details.

Terminology and protocol notes. We use *harmful-like* for endpoint-observable or judge-flagged harmful-like behavior in sanitized summaries, not for raw harmful prompt content. *Endpoint-effective*, *endpoint redirection*, and *correction* refer to judged shifts toward non-harmful endpoint behavior under the stated audit protocol.

Regime	Count	Role	Endpoint-observability use
Native baseline	1368	Model-local H/R/C/B calibration; z-score normalization; reference surfaces	No endpoint fitting
Discovery / attack-wrapped	up to 1082, depending on model-specific audit	Functional morphology, sparse fitting, source ranking	No held-out endpoint fitting
JBB held-out	555 per Qwen2.5 scale	Endpoint-observable boundary probe and posterior intervention audit	harmful-like: 353/555 (72B), 406/555 (32B), 451/555 (7B)
StrongREJECT	313/200/200 by scale	Saturation check	harmful-like: 1/313 (72B), 1/200 (32B), 4/200 (7B); not used as the main endpoint window when saturated

Table 3: Prompt regimes and endpoint observability. Native and discovery rows define model-local coordinates and source rankings; held-out JailbreakBench is reserved for posterior endpoint-observable audits. StrongREJECT is used as a saturation reference for Qwen2.5 rather than as the main endpoint-transition window.

H/F movement denotes changes in fitted harmfulness/refusal functional scores; reverse rows report signed H/F increases, whereas suppression rows report ΔF reductions under their own protocol. Generation-phase slots are fixed cached checkpoints over prompt/continuation phases rather than learned time variables; generated-late metrics are computed only over declared late cells in that schedule. IO-fitness denotes judge-scored fluency, coherence, and context preservation, with GPT/MiMo-style judges primary and malformed auxiliary rows treated as uncertainty. These are diagnostic metrics, not endpoint-control guarantees.

5 Results

The results are organized by the audit questions that SAGE is designed to answer: whether a benchmark exposes endpoint transitions, whether formula-selected source windows move the intended internal geometry, and whether the readout coordinates are stable under token/background perturbations.

RQ1: Do endpoint benchmarks expose safety-boundary transitions? SAGE first asks whether a benchmark supplies a measurable endpoint boundary. If a benchmark is saturated into safe outcomes, a probe can move internal scores while endpoint metrics remain uninformative. Table 3 shows that held-out JailbreakBench provides endpoint-observable Qwen2.5 boundary cases, while StrongREJECT is saturated in this family. We therefore use JailbreakBench for posterior endpoint-observable audits and treat StrongREJECT as a saturation reference.

RQ2: Can SAGE identify source windows that produce targeted internal movement? Reverse induction is a diagnostic stress test of functional movement, not a jailbreak method and not an endpoint-conversion claim. Table 4 shows that targeted reverse conditions produce much larger harmfulness/refusal functional movement than random controls across Qwen2.5 scales. For example, in 32B, targeted H/F movement reaches 11.78 and 10.65 on the two target conditions, while random movement stays near zero. These rows are functional movement evidence only; endpoint conversion is evaluated separately and is not inferred from H/F movement.

RQ3: Are the functional coordinates stable under readout perturbations? Table 5 reports a Qwen2.5-7B cached-feature ablation over 50 regular-benign native prompts and 50 harmful-like or boundary prompts. The check perturbs concept-token and background choices at the aggregate-feature level and includes a random-concept negative control. The results support ranking stability under cached proxy perturbations while preserving the intended claim boundary: they are not a substitute for full residual-to-token recomputation for every token-set variant.

RQ4: Can a source-ranked probe improve endpoints without measured refusal damage? Among the endpoint and preservation audits, the cleanest appendix result is the Qwen2.5-7B V-risk source row: it obtains an 11.2% correction rate with 0.0% measured R-damage. We promote this compact result because it supports the preservation side of the four-axis audit: SAGE is not only measuring movement, but can identify source families whose endpoint shifts are not purchased by the measured refusal-damage axis. We do not promote the higher-

Model	Condition	n	H/F increase	Endpoint audit status
14B	H/V-risk	10	9.59 / 8.27	aggregate functional rows; staged endpoint audit separate
14B	Morphology	10	10.31 / 20.53	aggregate functional rows; staged endpoint audit separate
14B	Random	10	0.01 / -0.23	aggregate functional rows; staged endpoint audit separate
32B	H/V-risk	10	11.78 / 10.65	aggregate functional rows; staged endpoint audit separate
32B	Morphology	10	7.34 / 13.64	aggregate functional rows; staged endpoint audit separate
32B	Random	10	0.23 / 0.57	aggregate functional rows; staged endpoint audit separate
72B	H/V-risk	3	7.32 / 2.93	small-sample functional rows; staged endpoint audit separate
72B	Morphology	3	3.67 / 3.70	small-sample functional rows; staged endpoint audit separate
72B	Random	3	0.01 / -0.09	small-sample functional rows; staged endpoint audit separate
7B	H/V-risk	10	9.71 / 6.74	aggregate functional rows; staged endpoint audit separate
7B	Morphology	10	6.37 / 8.67	aggregate functional rows; staged endpoint audit separate
7B	Random	10	-0.04 / 0.11	aggregate functional rows; staged endpoint audit separate

Table 4: Qwen cross-scale reverse functional movement summary at the implementation saturation level ($\alpha = 12$). The H/F column reports H increase / F increase. The table compares targeted versus random internal movement across scales; endpoint conversion is evaluated separately and is not inferred from these functional rows.

Table 5: Minimal cached readout robustness ablation on Qwen2.5-7B prompt-level functional features. These are proxy perturbations over cached aggregate features, not full residual-to-token recomputation.

Ablation	Surface corr.	F Spearman	V-risk corr.	Top-10 overlap	Source-rank corr.
Original T_0, no perturbation	1.000	1.000	1.000	1.000	1.000
20% concept-token drop proxy	0.982	0.995	0.998	0.966	0.989
40% concept-token drop proxy	0.946	0.984	0.994	0.934	0.968
Synonym/category-neighbor proxy	0.992	0.991	0.999	1.000	0.995
Frequency-matched T_0 proxy	0.999	1.000	0.999	1.000	0.987
Random frequent-neutral T_0 proxy	0.994	0.999	0.996	0.900	0.971
Random-concept negative control	0.010	0.091	0.106	0.100	0.063

correction 72B H-genlate frontier row because its 53.2% R-damage makes it a damage-tradeoff case rather than a clean positive result.

RQ5: Does source-window functional movement extend beyond Qwen2.5? Appendix stress tests on Llama-3.1-8B, Mistral-Nemo-12B, and Gemma-2-9B show the same targeted-versus-random functional-movement separation outside the main Qwen2.5 family. These rows strengthen the contribution as a diagnostic atlas rather than a Qwen-specific artifact. They are included as functional stress-test evidence only; endpoint conversion is still audited separately and is not inferred from H/F movement.

Endpoint movement, quality, and preservation are not treated as implicit consequences of internal movement. They are separately audited under the four-axis decomposition in Table 2, with additional endpoint-redirection, quality, preservation, fixed-schedule baseline, and non-Qwen stress-test diagnostics reported in Appendices E, G, and N.

6 Conclusion

SAGE reframes safety-boundary analysis as a prompt-wise functional-geometry audit. Rather than asking only whether an endpoint label flips, it first models internal safety geometry and then measures which fitted geometry moved before interpreting endpoint behavior separately. Across Qwen2.5 scales, formula-selected source windows induce substantially larger harmfulness/refusal functional movement than matched random controls, and endpoint-observable benchmark regimes differ sharply between JailbreakBench and StrongREJECT. The contribution is the functional atlas

Promoted appendix audit	Positive result moved to main text	Contribution strengthened
Preservation-positive source audit	Qwen2.5-7B V-risk reaches 11.2% correction with 0.0% measured R-damage.	Shows that source-window ranking can expose a clean endpoint/preservation operating point, not only internal movement.
Non-Qwen functional stress test	Targeted vs. random H/F movement: Llama-3.1-8B 43.739 vs. -3.203; Mistral-Nemo-12B 19.057 vs. 0.356; Gemma-2-9B 9.697 vs. 0.051. Targeted strong rates are 98.3%, 94.2%, and 46.7%, while random strong rates are 3.3%, 11.7%, and 0.0%.	Supports architecture-local generality of the atlas-selected source directions while preserving the paper's boundary that functional movement is not endpoint conversion.

Table 6: Appendix experiments promoted to the main text. We include only positive, contribution-strengthening rows: a preservation-positive Qwen2.5-7B source result and non-Qwen targeted-versus-random functional-movement stress tests. Higher-damage or underpowered endpoint-control rows remain in the appendix.

and audit protocol: SAGE is not a discovered circuit, a proof of causal mechanism, or a deployable defense. It gives a reporting protocol for activation-space audits in which internal movement is measured directly instead of being collapsed into a single endpoint score.

Limitations

SAGE is a diagnostic audit protocol, not a deployable defense. Source windows are operational coordinates; endpoint-rate rows are underpowered; non-Qwen evidence is staged; stronger state-dependent baselines remain future work.

Ethical considerations

This work studies safety-boundary auditing. Public artifacts exclude raw harmful prompts, raw model generations, private judge logs, unsafe generated text, credentials, and endpoint access details. Released materials contain sanitized figures, aggregate metrics, labels, hashes, and package manifests for verification while avoiding publication of operational harmful content. The public GitHub release will prioritize safe reproducibility: scripts, sanitized manifests, aggregate metrics, hashes, figures, environment metadata, and experiment details will be provided, while operationally harmful raw content remains excluded.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in Language Models Is Mediated by a Single Direction. In *Advances in Neural Information Processing Systems*.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting Latent Predictions from Transformers with the Tuned Lens. arXiv preprint arXiv:2303.08112.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. 2024. JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse Autoencoders Find Highly Interpretable Features in Language Models. arXiv preprint arXiv:2309.08600.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. 2024. A StrongREJECT for Empty Jailbreaks. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering Language Models With Activation Engineering. arXiv preprint arXiv:2308.10248.

Fred Zhang and Neel Nanda. 2023. Towards Best Practices of Activation Patching in Language Models:

Metrics and Methods. arXiv preprint arXiv:2309.16042.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. 2023. Representation Engineering: A Top-Down Approach to AI Transparency. arXiv preprint arXiv:2310.01405.

SAGE Supplementary Material

A Readout Details

The H/R/C/B readouts are concept-token contrasts against a background set. For model m, prompt p, residual state $h_{\ell,t}(p)$, and selected output-readout row e_u, the token score is

$$z_u(\ell, t; p) = \langle h_{\ell,t}(p), e_u \rangle + b_u,$$

with b_u omitted when the architecture has no output bias. The implementation uses model-local vocabulary readouts/logit-lens scores rather than a supervised endpoint probe. For field a, the raw readout is

$$r_a(\ell, t; p) = \operatorname*{logsumexp}_{u \in T_a} z_u(\ell, t; p) - \operatorname*{logsumexp}_{u \in T_0} z_u(\ell, t; p).$$

Native prompts define model-local means and variances for every field, layer, and phase, producing calibrated $\tilde{r}_a$ values. T_{H}, T_{R}, T_{C}, and T_{B} are built from semantic seed terms for harmfulness-like boundary content, refusal/safety, procedural compliance, and benign continuation. Terms are tokenized with the target model tokenizer and filtered to remove special tokens, punctuation/function fragments, and exact overlaps across concept groups. T_0 is a neutral background set used to subtract generic readout mass; background variants are audited in Appendix K. Public appendix material reports sanitized construction rules and omits raw harmful prompts or generations. Raw traces are retained only in a private academic verification package outside the public manuscript.

B Sparse Fitting Diagnostics

Sparse equations are fit on model-local functional features. Diagnostic controls include length-only baselines, prompt-length baselines, residualized analyses where available, label-shuffle checks, and coefficient-stability estimates. High raw AUROC in 7B/32B is interpreted cautiously because length baselines are strong.

C Protocol Clarifications and Existing Evidence

This appendix consolidates protocol details that are otherwise distributed across the method, experiment setup, and audit tables. It does not introduce new experiments; it states what the existing evidence supports and where the claim is intentionally bounded.

Generation-phase indexing. The phase variable t is a discrete cached readout schedule over prompt/continuation checkpoints, not a learned continuous-time variable. Layer–phase cells are therefore operational coordinates. Generated-late metrics are computed only over the declared late cells in this schedule, so they should be read as schedule-conditioned diagnostics rather than response-length-invariant quantities.

Energy/risk functional scope. The risk score combines late harmfulness energy, H–R excess, persistence, coupling, finite-difference movement, and preservation terms. Its sparse coefficients are model-local and fit on discovery/source-family splits. SAGE uses this scalar to rank source windows and design audits; it is not an endpoint classifier, a universal safety score, or a proof of stability. Term-level ablations and coefficient transfer tests remain outside the current evidence, so the main text reports functional movement and endpoint behavior separately.

Concept-token readout scope. The H/R/C/B fields use concept-token and neutral-background contrasts. This makes the atlas lightweight and model-local, but it also exposes the readout to token-set, tokenizer, frequency, language, and template choices. Table 5 reports cached aggregate perturbation checks and a random-concept negative control; these checks support ranking stability under proxy perturbations but do not replace full residual-to-token recomputation.

Judge and IO-fitness scope. IO-fitness is a judge-side quality axis covering fluency, coherence, and preservation of task-relevant context. Existing GPT/MiMo-style judgments are primary where available; DeepSeek is auxiliary only where parsed outputs are reliable. Presentation-perturbation judg-

ing was not run, so Table 12 reports existing multi-judge and parse robustness rather than perturbation stability.

D Intervention and Control Details

Each source window is defined by source family, ranking rule, k, α, and preservation constraints. Random-equal-norm controls match perturbation norm, while off-cell same-direction controls test source locality. Held-out endpoints are used only for evaluation.

E Endpoint, Quality, and Preservation Audit Details

Forward endpoint probes. The adaptive forward audit asks whether source-ranked perturbations produce endpoint movement under an upper-bound escalation schedule, but these rows should be interpreted as exploratory endpoint diagnostics rather than an intervention benchmark. Table 8 reports single-condition adaptive probes. The best single endpoint-effective condition is the adaptive global-direction probe at 7/31 prompts. Source-shaped conditions reach 4/31, while the adaptive random equal-norm control reaches 3/31. These endpoint-rate differences are small and do not establish endpoint-control superiority; they are retained as diagnostic evidence that endpoint behavior should be audited separately from internal functional movement.

Endpoint redirection and output quality. Endpoint non-harmfulness is not sufficient evidence of a good intervention. A redirected answer can be useful, over-refusing, degraded, or uncertain. Table 9 reports the adaptive candidate audit: 21/24 candidate rows have endpoint-nonharmful majority labels, while only 11/24 have high input-output fitness majority labels. This appendix result motivates keeping endpoint behavior and output quality as separate audit axes.

Correction and preservation cost. Table 10 summarizes the source frontier. The 72B H-genlate source has the largest correction rate, 9/47 or 19.1%, but also has 25/47 or 53.2% R-damage. This diagnostic result illustrates why SAGE reports correction rate and preservation cost as separate axes rather than treating endpoint correction as a complete success criterion.

F Repeatability and Safety Audits

The 72B repeatability audit repeats single-shot shifts and finds no stable 7/10 endpoint-control case. Public safety audits report category-level unsafe unlock results only. Private raw prompts, outputs, and judge logs are excluded from public materials.

G Compute and Baseline Feasibility

Atlas construction requires forward readouts over the selected layer and phase schedule. Intervention evaluation requires patched generation and judge adjudication. The appendix distinguishes the adaptive global-direction probe in Table 8 from the fixed-schedule norm-matched global-direction baseline described here. The fixed-schedule baseline is intentionally simple: it uses one model-local H/R/C/B readout-centroid direction, applies the same fixed layer/phase schedule for every prompt, and is not tuned on held-out correction outcomes. Direct SVF/conceptor/GSAE-style comparisons are not included in this submission because they require additional state-dependent implementations and parity checks rather than a single matched global vector.

Compute environment and setup. The experiments used approximately 120 GPU-hours on a dual-GPU cloud server with two Pro6000-class GPUs. Because runs were assembled from staged cached-readout, patched-generation, and post-hoc judging pipelines, this should be read as an approximate aggregate compute budget rather than a normalized per-experiment estimate. Model execution uses model-local tokenizers and inference configurations for the Qwen2.5 scale family and the Llama-3.1-8B, Mistral-Nemo-12B, and Gemma-2-9B stress tests. Core setup details reported in the manuscript and supplement include the native calibration pool, JBB held-out boundary set, StrongREJECT saturation audit, source-window definitions, top-k windows, alpha schedules, adaptive escalation/stopping rules, per-cell step caps, random equal-norm controls, off-cell controls, and staged-judge candidate rules. Implementation scripts and package metadata are included in the reproducibility packages where available; exact package versions refer to dependency and runtime versions such as PyTorch, Transformers, tokenizers, accelerate, CUDA/runtime libraries, judge-client libraries, and model checkpoint revisions.

Audit question	Existing evidence used in this paper	Claim boundary
Endpoint efficacy	Table 8 reports single-condition adaptive probes; best single condition is 7/31 and source-shaped probes reach 4/31.	Exploratory endpoint movement only; no statistically significant endpoint-control or defense claim.
Functional movement	Table 4 compares Qwen2.5 targeted reverse probes with random controls; Appendix N reports staged non-Qwen stress tests.	Functional movement evidence, not harmful endpoint conversion evidence.
Readout robustness	Table 5 reports cached token/background perturbation proxies and a random-concept negative control.	Supports proxy stability only; no full residual-to-token recomputation or multilingual validation claim.
Judge reliability	Tables 9 and 12 separate endpoint labels, IO-fitness, parse reliability, and staged non-Qwen judgments.	Malformed or unresolved rows are uncertainty; no full human annotation or presentation-perturbation robustness claim.
Baselines	Random equal-norm, off-cell same-direction, adaptive global-direction probe, and a fixed-schedule norm-matched global baseline are reported.	No superiority claim over state-dependent refusal-vector, SAE, SVF, conceptor, or GSAE-style baselines.
Source windows	Source windows are selected from fitted layer–phase functionals and compared against locality controls.	Operational source windows, not discovered circuits or causal mechanism proof.

Table 7: Existing evidence and claim boundaries for SAGE audits. The table makes the diagnostic scope explicit without adding new experiments.

Condition	Eff.	Rate	Clean	Resistant	ΔH	ΔF
Anti-H source probe	4	12.9%	4	27	5.19	2.97
Adaptive global-direction probe	7	22.6%	7	24	3.18	2.97
Morphology-matching probe	4	12.9%	4	27	2.83	3.34
Off-cell same-direction control	4	12.9%	4	27	2.88	3.29
Peak-clamp probe	3	9.7%	3	28	2.90	3.32
Random equal-norm control	3	9.7%	3	28	0.06	0.08

Table 8: Qwen2.5-7B adaptive audit summary. Rows are single-condition probes under the same escalation schedule and stopping rule, not endpoint-control wins. Eff. denotes endpoint-effective rows. Endpoint-rate differences are descriptive and exploratory rather than significance claims.

Recorded runtime versions. The live experiment environment probe records Miniconda Python `3.12.3`, pip `24.0`, PyTorch `2.8.0+cu128`, CUDA availability as `True`, CUDA build/runtime `12.8`, cuDNN `91002`, NVIDIA driver `580.82.09`, and two `NVIDIA RTX PRO 6000 Blackwell Server Edition` GPUs with 97,887 MiB reported memory per GPU. Core Python package versions are: Transformers `4.57.6`, tokenizers `0.22.2`, accelerate `1.13.0`, NumPy `2.3.2`, pandas `3.0.3`, SciPy `1.17.1`, scikit-learn `1.8.0`, matplotlib `3.10.5`, seaborn `0.13.2`, datasets `4.8.5`, pyarrow `24.0.0`, safetensors `0.7.0`, sentencepiece `0.2.1`, protobuf `6.31.1`, tqdm `4.67.3`, triton `3.4.0`, requests `2.34.0`, httpx `0.28.1`, PyYAML `6.0.2`, Pillow `11.3.0`, and ModelScope `1.36.3`. The probed runtime did not include vLLM, FlashAttention, bitsandbytes, PEFT, OpenAI client SDK, or einops. Cached model configuration metadata was available for the Qwen2.5 family, Llama-3.1-8B-Instruct, and Gemma-2-9B-it; exact upstream model revision hashes were not present in the public sanitized metadata and are therefore not inferred.

H Responsible Research Checklist Support

Artifacts and intended use. Public artifacts are intended for research evaluation and paper verification only. The SAGE public release is MIT-licensed. It contains sanitized aggregate metrics, endpoint labels, hashes, figures, and package manifests. Raw harmful prompts, raw model outputs, private judge logs, and unsafe generations are excluded. Upstream benchmark and model artifacts remain subject to their respective public or research access terms, while the released SAGE code/tables/manifests use the MIT license. Some legacy package and file identifiers retain the historical `dynaatlas_` prefix for reproducibility and checksum continuity.

Metric	Value	Status
Judge-score rows	144	current audit subset
Output rows	48	current audit subset
Post-intervention rows	24	current audit subset
Endpoint-nonharmful majority	21/24	current audit subset
High-IO majority	11/24	current audit subset
DeepSeek parse success	low	malformed rows treated as uncertainty
Primary judge policy	GPT and MiMo primary	DeepSeek auxiliary where parseable

Table 9: Tri-judge status for the adaptive candidate audit subset. Endpoint-nonharmful majority is 21/24 and high-IO majority is 11/24. GPT and MiMo are primary; DeepSeek is auxiliary only where parsed outputs are reliable, and malformed rows are treated as uncertainty rather than positive evidence.

Model	Best source	Correction rate	R-damage rate
72B	H_genlate	19.1%	53.2%
7B	V-risk	11.2%	0.0%
32B	R-preserving	7.5%	12.5%

Table 10: Best conservative source family by scale. The 72B H-genlate row has the largest correction rate but also substantial R-damage (53.2%), so it is an effect–damage frontier point rather than a clean intervention.

Data and counts. The main text reports the 1368-prompt native baseline manifest, 555 held-out JBB prompts, and StrongREJECT saturation analysis. It also identifies the model families used for Qwen2.5 scale experiments and non-Qwen stress tests. The public package table below provides artifact IDs and SHA-tracked sanitized packages without machine-local paths.

Judging and annotation. No crowdworkers or human-subject participants were recruited. Endpoint and IO-fitness labels use LLM-assisted judges and author-side audits; the paper does not claim human annotation. GPT/MiMo-style judges are primary where available, DeepSeek is auxiliary only where parsed outputs are reliable, and unresolved rows remain uncertainty rather than positive evidence.

I Fixed-Schedule Global-Direction Baseline Details

This fixed-schedule norm-matched global-direction baseline is distinct from the adaptive global-direction probe reported in Table 8. The Table 8 probe uses the adaptive strong-gate schedule and stopping rule on the 31-prompt adaptive audit subset, whereas this appendix baseline applies a fixed layer/phase schedule on the 80-prompt harmful-branch baseline set. It is included to test whether SAGE merely benefits from applying a generic harmful-to-safe steering vector. The direction is computed as a background-centered readout contrast, $\mathrm{norm}((R+B)/2-(H+C)/2)$, and is applied to a fixed eight-slot layer/phase schedule. The perturbation norm is matched per sample to the Qwen2.5-7B SAGE V-risk $k=8, \alpha=0.65$ reference condition. The baseline is evaluated on the same 80 harmful-branch prompts and a 30-prompt preservation mini-check. It obtains 0/80 harmful-to-non-harmful corrections, 0/80 non-refusal-safe shifts, 26/80 degraded or uncertain outputs, mean $\Delta F = 26.27$, and 16/80 R-damage cases. The preservation mini-check observes 0/30 unsafe unlocks. These results are public only as hashes, categories, and aggregate metrics; raw prompts and outputs remain in private quarantine and are excluded from the public package.

J Formal Functional Definitions

Table 11 summarizes the operational functionals used in the atlas. Full public dictionaries and implementation paths are included in the sanitized formalization package.

K Minimal Robustness Ablations

Readout robustness. Table 5 reports a small Qwen2.5-7B cached-feature ablation over 50 regular-benign native prompts and 50 harmful-like or boundary prompts. The check perturbs concept-token and background choices at the aggregate-feature level: 20% and 40% concept-token drop proxies, synonym/category-neighbor proxies, frequency-matched and random frequent-neutral background proxies, and a random-concept negative control. Because this is a cached-feature proxy, it tests ranking stability under aggregate perturbations; it is not a full residual-to-token recomputation for every token-set variant.

Judge robustness. Table 12 summarizes existing sanitized judged rows without introducing raw outputs into the public artifact. The audit compares available judge labels and parse success across

Term	Operational definition	Domain	Interpretation
H/R/C/B readout	$r_a(\ell, t; p) = \mathrm{logsumexp}_{u \in T_a} z_u - \mathrm{logsumexp}_{u \in T_0} z_u$; $\tilde{r}_a = (r_a - \mu^{native}_{a,\ell,t})/(\sigma^{native}_{a,\ell,t} + \epsilon)$.	(p, ℓ, t)	Model-local calibrated field coordinate.
$F(p)$	$F(p) = E^{late}_H(p) - E^{late}_R(p)$, or the same H-minus-R margin over the declared source window Ω.	prompt trajectory	Basin-proximity surrogate, not endpoint proof.
ΔF	$F_{\mathrm{base}}(p) - F_{\mathrm{patched}}(p)$ for suppression rows; reverse rows report signed H/F increases separately.	paired intervention row	Functional movement in the intended margin.
ΔR	$E^{late,z}_{R,\mathrm{patched}}(p) - E^{late,z}_{R,\mathrm{base}}(p)$.	paired intervention row	Refusal/safety counterforce movement; negative values may indicate R damage.
$E^{late}_{H/R}$	$\sum_{(\ell,t) \in \Omega_{\mathrm{late}}} \max(\tilde{r}_{H/R}(\ell, t; p), 0)^2$, with Ω_{late} denoting generated-late cells.	generated-late cells	Late response-mode persistence.
Ridge/persistence	High-readout connected or top-quantile layer/phase regions; persistence counts adjacent depth/phase support above threshold.	layer-phase surface	Source morphology and late ridge stability.
Coupling	Correlation, signed product, or co-activation summary of calibrated H/R fields over the declared cells.	layer-phase cells	Boundary interaction between harmful and refusal fields.
Velocity/acceleration	First and second finite differences of calibrated fields over normalized depth or generation phase.	adjacent cells	Direction and sharpness of functional movement.
V-risk	Sparse weighted combination of calibrated H/R/C/B functionals fit on discovery split and evaluated post-hoc; coefficients are model-local.	source-ranking domain	High-sensitivity risk source, not an endpoint classifier.
Entropy/P-H concentration	Entropy and concentration of positive H-field mass over cells, e.g. normalized $-\sum_i q_i \log q_i$ with $q_i \propto \max(\tilde{r}_H(c_i), 0)$.	source morphology	Whether H mass is diffuse or concentrated.

Table 11: Concise functional dictionary. These definitions are operational diagnostics rather than PDE, Lyapunov, or circuit-localization theorems.

adaptive rows and staged non-Qwen candidates. GPT/MiMo are primary for existing judged rows, while DeepSeek is auxiliary only where parsed outputs are reliable; low parse reliability on the adaptive subset means unresolved or malformed rows are treated as uncertainty rather than positive evidence. A private hash-only perturbation manifest was prepared for 20 Llama candidate outputs under original, bullet/plain, harmless-wrapper, and whitespace/style variants, but the presentation-perturbation robustness rerun was not performed because local MiMo/DeepSeek judge credentials were unavailable in the Windows environment. This credential limitation applies only to that planned rerun and does not invalidate the existing sanitized judged rows. The paper therefore does not claim presentation-perturbation stability. Endpoint-nonharmful labels are stable for non-Qwen staged candidates, while exact safety-category agreement and IO-fitness are judge-sensitive, especially in the adaptive subset.

Table 12: Minimal judge robustness summary over existing sanitized judged rows. A private hash-only manifest was prepared for presentation perturbations over 20 Llama candidate outputs, but the external perturbation rerun was not performed because no local judge credentials were configured. This table therefore reports multi-judge and parse robustness only.

Set	n	Safety cat.	Endpoint	IO$\geq$4	Status
Adaptive after: MiMo vs GPT/manual	24	0.125	0.333	0.333	Existing adaptive subset; DeepSeek auxiliary only where parsing is reliable.
Llama staged candidates	75	0.667	1.000	0.467	Existing staged rows; perturbation rerun not performed.
Gemma-2 staged candidates	75	0.907	1.000	0.427	Existing staged rows; perturbation rerun not performed.
Mistral-Nemo staged candidates	75	1.000	1.000	0.880	Existing staged rows; perturbation rerun not performed.

L Source Direction and Preservation Manifests

Table 13 gives the source-family construction used in the intervention audit. Held-out endpoint outcomes are not used to select source families, concrete cells, k, α, or directions.

Preservation constraints. The intervention window is written as $K_{\text{window}} = (s, r_s, k, \alpha, \epsilon_R, \epsilon_{\text{KL}}, \epsilon_B)$. In the current implementation, ϵ_R is monitored operationally via $\Delta R_{\text{genlate}}$, R-damage flags, and preservation-branch audits. ϵ_{KL} and ϵ_B are monitored preservation axes rather than hard online constraints in this implementation. Therefore preservation claims are empirical and condition-specific; nonzero unsafe unlock blocks any deployable-defense or guaranteed-safe-intervention interpretation.

M Reproducibility and Public Artifact Boundary

This section supports reproducibility and artifact-boundary checklist items. Public releases contain sanitized tables, aggregate metrics, labels, hashes, figures, and manifests; they exclude raw prompts, raw model outputs, private judge logs, credentials, machine-local paths, and endpoint access details.

Source	Features/cells	Direction construction	Split usage	Preservation monitoring
H-genlate	Persistent late H ridges.	Per-cell descent/rebalance direction from source morphology against native/reference states.	native + discovery; held-out evaluation only.	R-damage and preservation branch.
V-risk	High continuous V-risk / H-minus-R sensitivity cells.	Risk-margin reduction direction normalized to fixed norm.	native + discovery; held-out evaluation only.	R-damage, degraded output, unsafe unlock.
R-preserving	H ridge cells passing R-preservation gates.	H contraction direction filtered to avoid local R suppression.	native + discovery; held-out evaluation only.	R-damage gate monitoring.
H-R excess	Cells with high harmful-minus-refusal excess.	Excess-margin reduction direction.	discovery ranking only.	R and degradation monitoring.
HR rebalance	Cells with unstable H/R coupling.	Rebalance direction over coupled H/R readouts.	discovery ranking only.	R and endpoint transition audit.
Global direction	Fixed layer/phase schedule, no source-local ranking.	Background-centered readout contrast $\mathrm{norm}((R+B)/2-(H+C)/2)$.	native/discovery summaries; held-out evaluation only.	matched norm, R damage, preservation branch.
Random/off-cell	Random equal-norm cells or off-source cells.	Norm-matched perturbation or same direction at non-ranked cells.	control only.	degradation and unsafe unlock audit.
SVF-lite	State-dependent global direction scaled by current margin.	Bounded margin-dependent scaling of a global direction; no source-local top-k.	baseline pilot only.	matched metrics and R damage.

Table 13: Source-family direction manifest. The table distinguishes source-local windows from global and control baselines.

Package	Artifact ID	Public package / SHA256	Public boundary
Formal/stat/repro	`dynaatlas_formal_stats_repro_patch_v1`	`dynaatlas_formal_stats_repro_patch_v1_public_sanitized_20260522_1127.tgz` `9d7258badf3e5ba0a33540e38c1e34b41c10ee3a93fa317e9124eddc8324a8d5`	safety scan PASS
Cross-arch + baselines	`dynaatlas_cross_arch_stronger_baselines_v1`	`dynaatlas_cross_arch_stronger_baselines_v1_public_sanitized_20260522_1147.tgz` `30e28f9367ad03368604bf240acbd915f41c497d25c8c036c5fe2a2270514c65`	safety scan PASS
Llama mini audit	`llama31_8b_mini_continuation_audit_v1`	`llama31_8b_mini_continuation_audit_v1_public_sanitized_20260522_1311.tgz` `29b1f6a35473d11536527e3a6efb0e5ad6bc84eb527e82129162f6f0504352a3`	safety scan PASS

Table 14: Public reproducibility packages. No raw prompts, raw model outputs, private judge logs, or unsafe generations are included. Public artifacts are identified by artifact ID, package name, and SHA rather than machine-local paths.

N Non-Qwen Reverse-Induction Stress Tests

This appendix reports staged reverse-induction stress tests on Llama-3.1-8B, Mistral-Nemo-12B, and Gemma-2-9B. They are appendix diagnostics rather than main results. The goal is to test whether targeted model-local reverse directions move atlas-defined H/F functionals more than random equal-norm controls, and whether endpoint conversion follows that movement. The experiments are not attack methods or jailbreak searches. Public artifacts contain only hashes, labels, aggregate metrics, and figures; raw prompt/output traces remain in the private academic verification package.

Protocol. Each model uses 20 safe or non-harmful prompts and five conditions: targeted H/V-risk reverse induction, targeted harmful-reference morphology matching, random equal-norm reverse perturbation, off-cell same-direction reverse perturbation, and a global harmful-direction reverse baseline when available. Each condition is evaluated at $\alpha \in \{3, 6, 12\}$, giving 300 non-baseline forward/intervention rows per model. All forward rows are first labeled by heuristic endpoint screening; 75 candidate rows per model are selected for staged judging. Candidate selection includes heuristic endpoint changes, large H/F movement, possible drift or degradation, and random/off-cell controls. Candidate judging uses MiMo2.5 as primary, private reading audit for all candidate rows, and DeepSeek V4 Flash only for disagreement, representative positive and negative cases, and per-condition spot checks.

Interpretation. The stress tests support a limited non-Qwen extension of the main claim. Llama gives the strongest functional-induction contrast, Mistral gives the cleanest endpoint-preservation stress test, and Gemma gives weaker targeted induction with more uncertain judged candidates. Together, they support functional/endpoint separation and architecture-local endpoint coupling beyond Qwen2.5 while remaining stress-test evidence rather than a finished cross-architecture replication claim.

Evidence boundary. These are staged-judged stress tests, not full exhaustive tri-judge audits. The 19/75 Llama uncertain/invalid rows (including 1 invalid row) and 24/75 Gemma uncertain or ambiguous candidate rows remain unresolved evidence. Heuristic endpoint labels are used only for candidate selection and may over-flag possible drift. DeepSeek parse reliability remains a qualification on the adaptive subset, and MiMo/private-reading/DeepSeek label disagreements are handled conservatively. Model-local harmful-reference morphology comes from available mini-audit or endpoint-observable pools and should be treated as a stress-test reference, not a fully matched discovery run.

Qwen cross-scale heuristic flags. For transparency, Table 18 preserves the reverse-specific candidate flag that was previously exposed in the main table. This flag is not a judge-confirmed endpoint label and is not used as evidence of harmful endpoint conversion.

Model	Targeted H/F	Random H/F	Targeted strong rate	Random strong rate
Llama-3.1-8B	43.739	-3.203	98.3%	3.3%
Mistral-Nemo-12B	19.057	0.356	94.2%	11.7%
Gemma-2-9B	9.697	0.051	46.7%	0.0%

Table 15: Targeted versus random reverse-induction movement in non-Qwen stress tests. H/F movement is $\Delta H_{\text{peak}} + \Delta F_{\text{peak}}$ over non-baseline forward rows. Targeted rows move internal H/F peaks more than random equal-norm controls, but endpoint conversion is not inferred from H/F movement alone.

Model	Candidate rows	Final harmful	Harmless / avoidance	Uncertain / invalid
Llama-3.1-8B	75	0/75	56/75	19/75 uncertain/invalid (incl. 1 invalid)
Mistral-Nemo-12B	75	0/75	75/75	0/75
Gemma-2-9B	75	0/75	51/75	24/75 uncertain; off-cell note

Table 16: Staged judge outcomes for non-Qwen stress tests. The judged candidate subsets do not confirm harmful endpoint conversion. Unresolved rows are retained as uncertainty rather than converted into positive evidence. These are staged candidate audits, not exhaustive full tri-judge audits.

Model	Public package	SHA256
Llama-3.1-8B	dynaatlas_llama31_8b_reverse_induction_stress_test_v1_public_sanitized_20260523_0756_staged.tgz	5721d187abe04558953321d03fef83ea771fd921a07b44ffb4be0fc81453feac
Mistral-Nemo-12B	dynaatlas_mistral_nemo_12b_reverse_induction_stress_test_v1_public_sanitized_20260523_1339_stagedfix.tgz	f1596d7aad9e4cc3d948428fe7c426b27140d4970eb3fa38fd1868259e183e93
Gemma-2-9B	dynaatlas_gemma2_9b_reverse_induction_stress_test_v1_public_sanitized_20260523_1339_stagedfix.tgz	9beed92959861ecb5297905b19cb234cbccfb399d3eede4894fdada899a05713

Table 17: Public sanitized packages for non-Qwen reverse-induction stress tests. Public materials include aggregate tables, hashes, labels, and figures only; raw prompt/output text is excluded.

Model	Condition	n	H increase	F increase	Heuristic reverse-risk flag	Degraded
14B	H/V-risk	10	9.59	8.27	1/10 (10.0%)	0/10 (0.0%)
14B	Morphology	10	10.31	20.53	1/10 (10.0%)	0/10 (0.0%)
14B	Random	10	0.01	-0.23	2/10 (20.0%)	0/10 (0.0%)
32B	H/V-risk	10	11.78	10.65	7/10 (70.0%)	0/10 (0.0%)
32B	Morphology	10	7.34	13.64	7/10 (70.0%)	0/10 (0.0%)
32B	Random	10	0.23	0.57	6/10 (60.0%)	0/10 (0.0%)
72B	H/V-risk	3	7.32	2.93	1/3 (33.3%)	0/3 (0.0%)
72B	Morphology	3	3.67	3.70	1/3 (33.3%)	0/3 (0.0%)
72B	Random	3	0.01	-0.09	1/3 (33.3%)	0/3 (0.0%)
7B	H/V-risk	10	9.71	6.74	6/10 (60.0%)	0/10 (0.0%)
7B	Morphology	10	6.37	8.67	7/10 (70.0%)	0/10 (0.0%)
7B	Random	10	-0.04	0.11	7/10 (70.0%)	0/10 (0.0%)

Table 18: Heuristic reverse-risk candidate flags, not endpoint labels. These counts come from the reverse-specific ‘reverse_induced_harmful_drift‘ flag and are used only to select or audit candidate rows. They should not be interpreted as judge-confirmed harmful endpoint drift.