Linear Access in LLM Representations: An Audit of Readout, Control, and Transfer

Anonymous ACL submission

Abstract

Linear methods are often used to justify claims about readout, steering, editing, and truth-related reliability in language models, but those claims require different evidence. The same linear signal can support hidden-state readout, weaken sharply as token-level control, and transfer only through richer supervised access families. We study that separation with Flash-Jacobian, an operator-level audit built from region-conditioned hidden-to-hidden surrogates fit on hidden-state populations. Across Qwen-3.5-4B, Llama-3.2-3B, and Phi-3-mini, three results recur. First, local operator fidelity forms a middle-layer corridor followed by a late collapse band; split, occupancy, and partition checks preserve that ordering as a repeatable descriptive anchor. Second, descriptive locality does not by itself yield clean semantic control: hidden-response edits are predicted accurately, but token-facing log-probability and family-level responses stay moderate-to-weak, and boundary-token effects often dominate. Third, truth-related access depends sharply on access family: raw-state linear and score-free geometry families stay weak, whereas selected hidden-state and probe-rich composite families recover model-sensitive signal. The contribution is an operation-level rule for linear evidence in LLMs: what supports readout, what supports only bounded intervention, and what requires richer access families for transfer.

1 Introduction

Linear methods are central to how current work reads, edits, and steers language models. Linear probes ask whether linguistic or factual variables are decodable from hidden states (Hewitt and Manning, 2019; Belinkov, 2022); representation engineering and activation steering add contrastive directions to change behavior (Zou et al., 2023; Turner et al., 2023; Rimsky et al., 2024); and knowledge-editing work uses low-rank or localized changes to internal maps (Meng et al., 2022; Sun et al., 2026). These results are often grouped under one heading, linear access, even though they answer different questions. Linearly reading a variable, controlling a behavior, and transferring a detector are different operations and require different evidence.

This paper studies that separation directly. Instead of treating linear access as one claim family, we ask what kind of claim a given linear result actually supports: hidden-state readout, bounded intervention, or transfer. This also distinguishes our contribution from recent non-identifiability critiques. Before asking whether a behaviorally effective steering vector is unique, we ask whether readout, control, and transfer were ever the same inferential move to begin with. All claims are scoped to the tested architectures and benchmarks. The question is therefore narrower than whether linear structure exists in LLMs. The issue is whether success at one operation is being promoted too quickly into a stronger one. A readable hidden-state variable need not become a usable control direction, and a within-model detector need not define a transferable or coordinate-free access rule. This is the practical sense in which we use *validity envelope*: the range over which a linear result still supports a stronger downstream reading.

We answer with Flash-Jacobian, an operator-level audit built around cluster-representative Jacobians (CRJs), our shorthand for region-conditioned local operator surrogates. Given hidden-state pairs from a model layer, Flash-Jacobian partitions the input space and fits local first-order hidden-to-hidden operators. These are population-level local operators, not exact pointwise Jacobians. The method contribution is therefore not another local map in isolation, but one common instrument for testing LRH-style evidence at three different operation levels.

The audit has three linked results. First, local operator fidelity follows a consistent depth profile across the tested models: a middle-layer corridor

has the strongest held-out fidelity, while late layers collapse. This pattern is not well explained by centroid distance, low cluster occupancy, or a narrow raw Euclidean k-means choice; a first SAE-feature partition check also fails to reveal a distinct learned-feature regime after occupancy balancing. We therefore use the depth pattern as a family-sensitive descriptive anchor for the audit, with gate/rank statistics as structural markers rather than as a complete mechanistic account. Second, semantically coherent local directions do not generally act as clean concept dials. In local parameter-edit audits, hidden-response deltas are almost perfectly linear, but tracked token probabilities are only moderate-to-weak; in direct interventions, boundary and punctuation suppression can dominate the desired semantic effect. Third, truth-related access is access-family dependent. Raw-state linear baselines and score-free CRJ geometry remain near chance in our HaluBench audits, while selected hidden-state probes and probe-rich composite families recover signal within and partly across models. A prompt-template audit further tightens the same boundary: with model, layer, casebook, and PASS/FAIL vocabulary held fixed, in-template hidden-state access remains real, but cross-template transfer degrades materially and the derived model-correctness target itself can move sharply under prompt rewriting.

The contribution sits in that shared evaluation frame. Local linearization, prompt-conditioned Jacobians, sparse Jacobian models, activation transport, truth probes, and steering vectors all have close prior or contemporaneous work (Hernandez et al., 2024; Golden, 2025; Farnik et al., 2025; Szablewski and Masiak, 2025; Marks and Tegmark, 2024; Liu et al., 2024). Our contribution is the audit design that connects these literatures: the same operator stack is pushed through readout, control, and transfer, so that descriptive linearity, intervention behavior, and truth-related access are evaluated under one shared contract rather than in separate paper-specific setups. We use *validity envelope* for that target: the range over which linear evidence that works for readout can also be promoted to control or transfer claims.

Contributions. We contribute (1) an inference-separation audit for LRH-style evidence, (2) a CRJ-based operator-regime map with a middle-layer corridor and late collapse band across three model families, and (3) downstream boundary results showing that hidden-response linearity does not generally become token-level semantic control and that truth-related access is access-family dependent rather than score-free geometry.

2 Related Work

Linear representation and probes. Formal and empirical work shows that linear structure can arise in representation learning and can encode meaningful variables (Park et al., 2024; Jiang et al., 2024; Hewitt and Manning, 2019). At the same time, probe methodology has long warned that decodability does not by itself identify model use or causal mechanism (Belinkov, 2022). We extend that caution to operator-level and intervention settings by measuring where stronger linear inferences remain supported and where they weaken.

Local operators and Jacobians. Our closest neighbors are local linear and operator-based analyses. Relation-decoding work shows that first-order approximations can explain some factual computations but not all relations (Hernandez et al., 2024); other work studies prompt-conditioned local maps, sparse computational structure, residual-stream transport, and parameter-update geometry for continual post-training (Golden, 2025; Nettasinghe and Joseph, 2026; Farnik et al., 2025; Szablewski and Masiak, 2025; Wang et al., 2026). The boundary is explicit: Flash-Jacobian does not claim local linearization or Jacobian analysis as new. It fits many population-level cluster-representative hidden-to-hidden operators and uses them to route the same evidence through held-out fidelity, semantic intervention, and truth-related access tests, rather than optimizing weight-update integration or continual merging.

Steering and truth-related access. Activation steering can change model behavior in practical settings (Zou et al., 2023; Turner et al., 2023; Rimsky et al., 2024), but recent analyses emphasize non-identifiability, off-target effects, and gaps between decodability and causality (Venkatesh and Kurapath, 2026; Taimeskhanov et al., 2026; Li et al., 2026; Makelov et al., 2024; Huang and Chang, 2025; Bas and Novak, 2025). Truth and hallucination work likewise supports internal signal while warning against a single universal truth axis: truth directions can be layer-, dataset-, instruction-, and context-dependent, and recent cone, context, and basin accounts are explicitly non-one-dimensional (Azaria and Mitchell, 2023; Marks and Tegmark,

2024; Liu et al., 2024; Yu et al., 2025; Adarsh et al., 2026; Cherukuri and Varshney, 2026; Ying et al., 2026). Our truth-related section is therefore an access-family audit aimed at signal source rather than detector ranking.

Concurrent related paper. A concurrent anonymized submission studies post-deployment unlearning under a different audit target: target suppression together with semantic-neighbor preservation, low-bit stability, extraction resistance, and bounded recurrence (Anonymous, 2026). The overlap is the audit stance rather than the empirical object. That paper evaluates operating points for unlearning after model edits, whereas our paper evaluates when linear evidence in hidden representations supports readout, control, and transfer claims.

3 Operator Audit Protocol

3.1 Cluster-Representative Operators

For a model layer, collect hidden-state input-output pairs $\{(x_i, y_i)\}_{i=1}^{N}$. Flash-Jacobian partitions the inputs into K regions and fits a local operator per region:

$$y \approx \bar{y}_k + J_k(x - \bar{x}_k), \qquad a(x) = k. \tag{1}$$

The population target is the conditional best linear predictor:

$$J_k^\star = \arg\min_J \mathbb{E}\big[\|Y - \bar{y}_k - J(X - \bar{x}_k)\|_2^2 \mid a(X) = k\big]. \tag{2}$$

We call this object a cluster-representative Jacobian because it plays the role of a first-order local input-output map inside the audit. It is *not* an exact pointwise differential Jacobian. For attention-containing blocks, shrinking a token-space cluster does not remove context marginalization; the exact-differential reading is most literal for position-wise submodules such as MLPs. All main claims therefore use held-out surrogate fidelity and downstream audits rather than assuming exact differential recovery. The separate exact-JVP calibration supports this boundary rather than erasing it. Those rows are a validation target for the naming boundary, not the CRJ estimand used in the main audits: across a three-seed isolated-MLP check, literal JVP-diagonal fits still peak in the same corridor layers (Qwen layer 7 cosine 0.564 ± 0.004, Llama layer 14 cosine 0.471 ± 0.004, Phi layer 11 cosine 0.317 ± 0.003) while early and late anchors remain much weaker. The relevant comparison in these rows is to matched shuffled-null baselines and to the early/late anchors, not to an arbitrary absolute cosine threshold. The distinction matters throughout the paper. When a CRJ predicts hidden-state perturbation responses well, we treat that as evidence for a local operator model inside the audited region. We do not infer that the same direction will be a faithful token-facing intervention, a model-invariant semantic axis, or a probe-free truth detector. Those stronger claims are tested separately.

3.2 Scalable Fitting

Naively storing K dense $d \times d$ operators is prohibitive. Flash-Jacobian uses low-rank or diagonal operator families and stores $J_k = A_k B_k$, reducing storage from $O(Kd^2)$ to $O(Kdr)$ for rank r. The fitting pipeline combines three engineering choices: (i) tiled assignment, which avoids materializing an $N \times K$ distance matrix; (ii) sort-inverse accumulation, which groups points by assignment before per-cluster updates; (iii) Woodbury low-rank ridge updates, replacing dense $d \times d$ inverses with smaller cluster Gram inverses when $n_k \ll d$. On Qwen-3.5-4B, fitting $K = 2000$ local operators over about 100k tokens takes about 155 seconds on a single NVIDIA L4 for the fitting stage alone. This number is a cost envelope for the operator audit; downstream semantic and truth-related audits are costlier and are evaluated separately.

4 Experimental Design

We evaluate three primary open-model families—Qwen-3.5-4B, Llama-3.2-3B, and Phi-3-mini-4k-instruct—on three linked questions. Depth geometry uses WikiText-2 hidden states and held-out JVP-style fidelity measurements. Semantic intervention uses local directions and low-rank parameter-edit approximations across lexical families. Truth-related access uses HaluBench as the main anchor, with TruthfulQA and BoolQ as secondary checks reported in the appendix. All claims are scoped to these models, datasets, and access families. Unless otherwise noted, the main depth sweeps use matched whole-layer local fits under a shared held-out protocol; grouped counts, seeds, and multi-seed summaries are reported in the appendix uncertainty table. For depth, we ask whether the corridor/collapse pattern survives split hygiene, occupancy thresholds, partition alternatives, basis

Operation	Setup, metric, and key control	Main-text supported reading	What would require stronger evidence
Readout	Qwen/Llama/Phi-3 on WikiText-2; held-out JVP-style hidden response across depth; split, occupancy, partition, basis, and branch controls.	Middle-layer corridor and late collapse define a repeatable operator-surrogate regime map.	Exact pointwise Jacobian recovery and a universal best-layer rule do not follow from the same evidence.
Control	Local parameter-edit and intervention audits on lexical families; hidden-response, tracked log-probability, and continuation checks with matched controls.	Local hidden IO can be trust-region faithful, and some policy-conditioned next-token gains exist in responsive families.	Those gains do not become a clean global concept dial or a continuation-level steering law.
Transfer	HaluBench primary anchor, TruthfulQA/BoolQ secondary anchors; raw-state linear, selected hidden-state, and probe-rich families under grouped and held-out-model controls.	Truth-related signal is real in selected hidden-state and probe-rich access families.	The same evidence does not yield a score-free truth geometry or template-global factuality axis.

Table 1: Operation-level audit table. Each row states the supported main-text reading and the stronger reading that would require a different or stronger evidence block.

changes, and branch decompositions. For semantic control, we ask whether local hidden-response predictability survives as token-family specificity and continuation-level control. For truth-related access, we ask whether signal belongs to raw-state linear geometry, selected hidden-state probes, probe-rich composite families, or transferable feature families. The main text reports the central comparisons; the appendix records the broader control set, alternative cuts, and implementation details used to evaluate the same claims.

Operation-level audit questions. Table 1 states what the reported evidence supports for each operation and what would require a stronger evidence block.

5 Depth: A Corridor and a Collapse Band

Figure 1 shows the main depth result. Across the three tested models, local operator fidelity is low early, peaks in a middle corridor, and collapses late. The peak regions are Qwen layers 5 to 19 (peak cosine 0.56 at layer 7), Llama layers 4 to 15 (peak 0.50 at layer 14), and Phi-3 layers 8 to 15 (peak 0.27 at layer 11). This is an operator-fidelity result, not merely a static anisotropy result: it measures how well local hidden-to-hidden maps predict held-out perturbation responses. It serves as the descriptive anchor for the downstream sections rather than as a stand-alone claim of a universal depth law. Semantic comparisons and constructive local policies are tested mainly inside the corridor, while truth-related probes are deliberately allowed to choose layers outside it. That design prevents the depth result from being overread as a universal best-layer rule.

The late collapse is difficult to dismiss as a trivial clustering artifact. Residual mismatch tracks fidelity tightly, while centroid distance is weakly correlated with failure (below 0.30 in the main Qwen and Llama sweeps). Cluster occupancy is also not lower in the late bands; median train occupancy changes only marginally from corridor to late layers in the main K=64 sweeps. A stricter split-hygiene audit gives a second check: across 18 arms, corridor test cosine is 0.3330 under the official split and 0.3339 under a within-train held-out split, while pre-corridor and late-collapse means remain much lower. The mean train-test cosine gap is near zero, and occupancy thresholds from 1 to 30 samples per cluster barely move regime means. A partition-family audit gives a complementary boundary: on Qwen, raw, cosine, diagonal-whitened, PCA-whitened/Mahalanobis-proxy, gate-space, boundary-proxy, random-hash, and single-global controls keep the corridor near 0.481 and late collapse near 0.035. The original hard-coded Phi anchors are near floor in that stricter protocol, but a Phi-specific pinned-loader replication recovers layers 10/11/16/17 around 0.25–0.30 and keeps layers 27/31 low, while layer 12 remains split-sensitive. A Qwen layer-18 SAE-feature check adds a learned-feature stress test: balanced SAE reasoning/GSM8K/random-feature partitions remain nearly tied with raw k-means and null controls, though the public SAE is not trained for the MLP-input CRJ target used here. Additional appendix controls preserve the same read. Tested supervised semantic-lexical partitions do not create a distinct fidelity regime over raw, single-global, or random

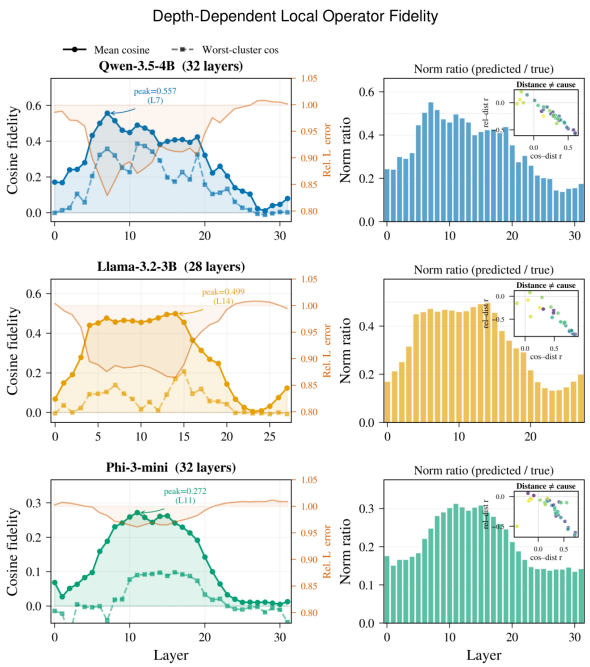

Figure 1: Depth-dependent local operator-surrogate fidelity across Qwen-3.5-4B, Llama-3.2-3B, and Phi-3-mini under the shared K=64 JVP-diagonal protocol. Cosine fidelity peaks in a middle-layer corridor and collapses late; residual mismatch and norm-ratio diagnostics track the same regime, while inset distance checks show that centroid distance does not explain the late-layer drop. Absolute levels and peak layers remain family-sensitive.

controls, while attention-state and attention–MLP-concat proxies are mixed and only show a modest late-Llama lift. The final partition comparison makes that quantitative: attention–MLP-concat is competitive but not dominant (best in 3/12 model-layer-split cells, mean Δcosine about -0.002 versus the best alternative partition), while supervised semantic-lexical wins only 1/12 cells and averages about -0.0086. The result is partition-stable operator approximability rather than a semantically privileged cluster atlas. These occupancy statements concern the whole-layer CRJ depth sweep. They do not contradict the exact-JVP density sweep in the appendix, which studies isolated-MLP surrogate-to-differential calibration under a different estimand and a different resolution/occupancy tradeoff.

The mechanism evidence is narrower. Gate effective-rank and anisotropy statistics track the late collapse strongly in raw depth-wise correlations, but this evidence is depth-confounded and partial-correlation checks weaken the point in some families. Branch decompositions further show family dependence: Qwen and Llama have MLP-led middle corridors under the main branch-local controls, while other families can be attention-led or weaker. Additional causal manipulations support a narrower statement: sharpening the dominant branch direction damages corridor fidelity, but simple inverse manipulations do not reverse late collapse. The depth result is therefore a reusable regime map for the audit, while the exact collapse mechanism remains family-sensitive.

Basis controls give the same bounded reading. At the favorable Qwen middle-layer anchor, diagonal, whitened, and cosine variants stay in the same narrow fidelity band, whereas the late anchor is much more basis-sensitive and is partially recovered only by an SVD-aligned basis. Llama shows a similar corridor stability with a weaker late SVD recovery, while Phi-3 keeps the qualitative corridor/collapse ordering at a lower absolute fidelity floor. This matters for interpretation: the middle corridor is not just a lucky coordinate choice, but neither is late collapse a basis-invariant absolute failure. The local operator family therefore exposes a depth regime whose strength and late-layer basis sensitivity are model-family dependent.

6 Semantic Control: The Descriptive-to-Causal Gap

The semantic audit asks whether directions that are locally meaningful become clean causal concept dials. At the descriptive level, they often look promising. K-means regions at Qwen and Llama middle layers isolate coherent token families such as surname prefixes, function words, date fragments, and appearance or career verbs. Operator spectra are moderately anisotropic rather than purely rank-one, but leading directions can be interpretable. The audit therefore proceeds in stages: interpretable hidden-state structure, token-family versus boundary effects, and then within-cluster transfer plus prospective policy checks.

Direct intervention shows the limit. In a Qwen layer-12 family/relational cluster, adding the leading operator direction changes family-token probability by only $+1.39 \times 10^{-5}$, while boundary and punctuation mass decreases by -5.70×10^{-3}. The raw ratio overstates the effect because target base mass is tiny, but the qualitative mechanism is clear: the dominant response is boundary suppression, not clean family activation. In the matched appendix control for this cluster, a CAA-style mean-difference direction is anti-aligned with the target

Audit row	Rows	Mean cos.	Median cos.	Mean rel. L2
Hidden-response delta	144	0.9892	0.9962	0.1645
Tracked log-prob delta	144	0.5586	0.6303	1.0796

Table 2: Local parameter-edit audit. Hidden responses are highly predictable by local linearization, but token-facing log-probability responses are much weaker.

CRJ direction (cosine −0.446) and does not improve the target-family contrast. Additional family clusters and matched mean-difference controls do not turn this into a general steering law.

Table 2 gives the larger version of the same boundary. Across 288 rows spanning three models, four depth regimes, two module types, and five lexical families, hidden-response deltas are predicted almost perfectly by local linearization. The tracked log-probability deltas are only moderate-to-weak and have relative error above one. The model breakdown follows the same hierarchy: tracked-response cosine averages 0.7432 for Qwen, 0.5687 for Llama, and 0.3640 for Phi-3, while hidden-response cosine stays near 0.99.

Recovered edit-atlas shards add a locality sanity check without supporting a full mechanism claim. Same-cluster hidden-response transfer is much stronger than cross-cluster transfer across Qwen, Llama, and Phi-3, while tracked token-facing transfer keeps the same qualitative locality advantage but remains much weaker. The atlas therefore supports the local trust-region premise while reinforcing the access boundary: local hidden IO structure transfers within regions; token-facing semantic effects do not become robust global controls.

Prospective semantic checks keep the constructive reading narrow. Across 945 target-control contrasts, decisive negatives and mixed cases dominate; phrase-level support concentrates in selected Qwen policies and lexical families, and continuation-level hard hits remain absent. Thus Flash-Jacobian exposes local geometric leverage, but in the tested semantic suite that leverage remains next-token-local rather than a continuation-level steering law.

Aggregate token-family statistics prevent the semantic section from resting on one anecdote. Positive-token absolute changes are only modestly larger than boundary-token changes on average, and a retrospective predictor audit tightens the same point: in pooled non-random rows, the strongest descriptive correlation with control-adjusted magnitude is about $\rho = 0.62$, while the strongest same-scope correlation with specificity is only about $\rho = 0.21$. Local conditional policies do improve next-token specificity in selected responsive Qwen clusters, but the winning policy changes by cluster, a matched Llama comparison is much weaker, and continuation readouts remain soft-only. The five-seed corridor continuation audit leaves that boundary intact: across 350 aggregate rows, only 26 have nonzero continuation rate, the mean target rollout-positive rate is 0.0022, and the maximum target rollout-positive rate is 0.0833. Those nonzero rows concentrate in a narrow Llama `kinship/temporal` pocket under mean-difference policies, while the matched Qwen arm still yields no surviving continuation-positive row. The resulting claim is therefore not ''steering never works''; it is that descriptive cluster purity is an insufficient certificate of semantic causal specificity.

7 Truth-Related Access: An Access-Family Ladder

For truth-related access, the key distinction is across access routes. Tested raw-state linear probes on HaluBench remain near chance, with best-layer logistic AUC 0.466 for Qwen, 0.514 for Llama, and 0.474 for Phi-3, while selected hidden-state and probe-rich feature families recover materially stronger signal. That split, rather than any single detector row, is the central finding. A matched prompt-template audit sharpens it further: even with fixed casebook, layer, and PASS/FAIL output space, mean cross-template AUC still drops by about 0.02 to 0.23, and the same case can flip across templates. A stable-target reconstruction audit partly explains that sensitivity but does not erase it: majority/supermajority/unanimous target rebuilding improves many slices, yet the mean rescue stays modest ($\Delta \approx +0.0250$ overall) and benchmark-level rescue remains mixed.

Selected hidden-state probes recover stronger but benchmark-sensitive signal under grouped source-, pair-, and length-style holdouts. Representative grouped rows reach AUC 0.775 on Qwen HaluBench and 0.772 on Llama TruthfulQA, while BoolQ is weaker at about 0.63. Near-perfect within-model composite rows appear only when the full probe-rich feature stack is retained; we treat these rows as upper-bound supervised access tests, not as headline detector performance. The feature-leakage ablation is the guardrail: when only the direct scalar score columns are removed, `core_minus_direct`

Access family	Typical AUC	Interpretation
Raw-state linear	0.466 to 0.514	too weak and biased on HaluBench
Selected hidden-state probes	0.63 to 0.78 in representative grouped rows	real but benchmark-sensitive signal
Probe-rich composite stack	near-perfect within model when direct probe channels remain	upper-bound supervised access, probe-dependent
Score-free geometry / trajectory	around chance in grouped audits	not a truth-geometry oracle
Raw-coordinate leave-model-out	about 0.50	no direct cross-model coordinate transfer
Metadata-free held-out-model probe-rich families	strong for held-out Qwen/Phi-3; 0.77 to 0.86 for held-out Llama	transfer is family-sensitive and probe-rich

Table 3: Truth-related access is access-family dependent. High detector scores trace supervised access families rather than a score-free geometric truth direction.

falls to roughly chance in grouped settings, and geometry-only or trajectory-only families stay near chance. It is therefore a direct-score ablation, not a removal of every probe-derived summary.

Table 3 summarizes the ladder. The central conclusion is that different internal access families carry sharply different amounts of signal. Raw linear geometry and score-free CRJ summaries are weak for the tested truth/fail targets; selected hidden-state and probe-rich families are stronger; and transfer depends on feature family and alignment. Full pooled raw-coordinate leave-model-out evaluation collapses to chance, while stricter metadata-free held-out-model audits preserve selected probe-rich transfer but not pure geometry. That transfer split matters because it rules out two easy overreads at once. The positive rows are not just a within-model artifact, since some metadata-free held-out-model probe-rich families do survive. But the result also does not collapse into a universal truth geometry, because the raw-coordinate leave-model-out family and the geometry-only families still fail. The supported reading is therefore a family-sensitive transfer hierarchy rather than a model-global truth axis.

Two guardrails keep this narrower. The strictest case-level pooled audits hold out the same HaluBench row across model families and remove explicit metadata, yet the same hierarchy survives: probe-rich rows remain strong, while geometry-only and trajectory-only families remain near chance. The direct-output decomposition shows that no single scalar explains the strongest rows; bundled probe outputs are stronger than `prob_fail` alone, but they remain supervised probe-derived channels. A nonlinear frozen-template audit reaches the same boundary: selecting among linear, PCA-linear, RBF-kernel, random-forest, and extra-trees classifiers on source-template validation chooses nonlinear families in about 70% of rows, but frozen cross-template AUC still drops by 0.144 on average and only about 22% of cross-template rows remain at AUC $\geq$ 0.70. Even the model-matched SAE comparison is negative for the strongest truth-related surrogate. The truth-related contribution is therefore an access-family map that identifies where signal becomes available and where it remains weak.

8 Discussion

The main lesson is to evaluate linear evidence by operation rather than by method family alone. The central question is what kind of operation a linear result survives: readout, bounded control, or transfer. One operator-level audit separates three regimes that are often conflated: descriptive local linearity, causal semantic control, and transferable truth-related access. The novelty is the evaluation contract itself: one operator-level audit asks how far the same piece of linear evidence can be promoted across those operations (Park et al., 2024; Jiang et al., 2024; Ying et al., 2026).

The depth result anchors that protocol. The middle-layer corridor and late collapse band survive split, occupancy, basis, and partition-family stress tests, but remain family-sensitive and are not a universal peak-layer law. Supplementary controls keep the same boundary: supervised semantic-lexical regions do not become privileged regimes, attention–MLP-concat is only competitive, and a richer low-rank bridge does not rescue the weak cross-token or attention-branch rows. The semantic and truth-related sections then show how stronger inferences fail in different ways. Semantically coherent local regions support trust-region hidden-response prediction, yet token-facing specificity stays much weaker and continuation-level positives remain narrow and unstable. For truth-related access, raw-state linear geometry stays weak, selected hidden-state and probe-rich families recover sig-

nal, and score-free geometry still fails to become a transferable detector ingredient even after grouped, held-out-model, and prompt-template audits.

This is also why the paper is not just a bundle of familiar cautions. Prior literatures separately warn that probes may overread, steering vectors may be non-identifiable, and truth-related directions may be benchmark-sensitive. What the present audit adds is a matched comparison under one instrument: the same operator family can look strong as descriptive readout evidence, degrade sharply as token-facing control evidence, and recover only through richer supervised access families under transfer. That operation gap is the main object of the paper, not an incidental restatement of earlier warnings.

The appendix scale checks are consistent with the same bounded reading. Additional 7B/8B/9B-family follow-ups reproduce the corridor-versus-collapse pattern non-uniformly, and Falcon-7B-Instruct serves as a negative family control rather than another positive headline row. This does not justify a universal law over all frontier models, but it does show that the audit is not confined to one checkpoint family or one benchmark slice.

The synthesis is that Flash-Jacobian is most useful as a trust-region diagnostic and a boundary-finding audit. The corridor identifies where local surrogates are descriptively strongest without collapsing depth choice into one best-layer rule for every downstream task. Likewise, strong probe-rich detector families identify informative access routes without collapsing truth-related access into a template-global or model-global truth geometry. The contribution is therefore a map of which stronger inferences remain supported by a given kind of linear evidence.

9 Conclusion

We have presented Flash-Jacobian, a scalable operator-level audit for testing how far linear evidence can be promoted in LLMs. Across the tested architectures, the same stack yields three coordinated conclusions: local operator fidelity follows a bounded depth regime with a middle-layer corridor and late collapse band; descriptively coherent clusters do not by themselves become clean global concept dials; and truth-related access is access-family dependent, with weak raw-state geometry and materially stronger selected hidden-state and probe-rich families.

Taken together, these results show that linear interpretability is better evaluated as a sequence of operations than as a single evidential category. Local linearization is descriptively informative in a bounded regime, stronger control claims require much more than descriptive coherence, and downstream truth-related signals depend on richer, benchmark- and family-sensitive access families rather than on raw linear geometry alone. Final partition-family and low-rank bridge controls sharpen that picture. The contribution is an audit for deciding what a linear result supports at the levels of readout, control, and transfer.

Limitations

All empirical claims are limited to the tested models, datasets, benchmarks, and operator families. CRJs are population-level local linear predictors rather than exact pointwise differentials, so the paper studies surrogate fidelity and downstream audit behavior, not exact local derivatives for full transformer blocks. The depth section identifies a repeatable regime map, but not a complete causal mechanism for late collapse. The semantic evaluation is concentrated on lexical families and mostly next-token-local effects, so it does not cover broader open-ended generation settings. The strongest truth-related results come from supervised and probe-rich access families, and the benchmark targets themselves are task-specific rather than a single unified notion of truth. All main evaluations are English-centric and white-box. Broader language coverage, closed-model settings, and black-box audit variants remain outside the scope of this paper.

Ethical Considerations

Interpretability methods can help identify biases and failure modes by making internal mechanisms more legible. At the same time, high-resolution internal analyses could facilitate targeted model manipulation, including steering toward unsafe behaviors or treating benchmark-specific internal scores as authoritative truth judgments. White-box access to hidden states may also raise privacy concerns if internal representations are used to infer sensitive training- or prompt-derived information. The experiments require GPU computation, although the core operator-fitting stage is designed to run on modest hardware and most additional audits are bounded post-processing or targeted checks. AI assistants were used in a limited supporting role

for English expression polishing, routine coding assistance, lightweight scripting, figure drafting, and internal manuscript diagnostics. All substantive research decisions, theoretical claims, experiment design, result interpretation, and final manuscript wording were reviewed and approved by the authors, who take full responsibility for the paper. We present the methods as diagnostic tools for transparent research, not as deployment-ready detectors or control mechanisms, and encourage use alongside established safety, privacy, and evaluation practices.

References

Shivam Adarsh, Maria Maistro, and Christina Lioma. 2026. How context shapes truth: Geometric transformations of statement-level truth representations in LLMs. *Preprint*, arXiv:2601.06599.

Anonymous. 2026. Beyond forget scores: Auditing quantization-robust llm unlearning with q-rou. Under review.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Tetiana Bas and Krystian Novak. 2025. What can we actually steer? a multi-behavior study of activation control. *Preprint*, arXiv:2511.18284.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread. Web research report.

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

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucy Farnik, Tim Lawson, Conor Houghton, and Laurence Aitchison. 2025. Jacobian sparse autoencoders: Sparsify computations, not just activations. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 16178–16236. PMLR.

Nathan Godey, Éric de la Clergerie, and Benoît Sagot. 2024. Anisotropy is inherent to self-attention in transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 35–48.

James R. Golden. 2025. Equivalent linear mappings of large language models. *Transactions on Machine Learning Research*. Published October 2025.

Shahar Haim and Daniel C. McNamee. 2026. Depthwise emergence of prediction-centric geometry in large language models. *Preprint*, arXiv:2602.04931.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2024. Linearity of relation decoding in transformer language models. In *The Twelfth International Conference on Learning Representations*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Lianghuan Huang and Yingshan Chang. 2025. Causality
neq decodability, and vice versa: Lessons from interpreting counting vits. *Preprint*, arXiv:2510.09794.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*. ICLR 2024 poster.

Yibo Jiang, Goutham Rajendran, Pradeep Kumar Ravikumar, Bryon Aragam, and Victor Veitch. 2024. On the origins of linear representations in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 21879–21911. PMLR.

Yuxiao Li, Alina Fastowski, Efstratios Zaradoukas, Bardh Prenkaj, and Gjergji Kasneci. 2026. Analysing the safety pitfalls of steering vectors. *Preprint*, arXiv:2603.24543.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human

falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Junteng Liu, Shiqi Chen, Yu Cheng, and Junxian He. 2024. On the universal truthfulness hyperplane inside LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18199–18224, Miami, Florida, USA. Association for Computational Linguistics.

Wen Luo, Guangyue Peng, Wei Li, Shaohang Wei, Feifan Song, Liang Wang, Nan Yang, Xingxing Zhang, Jing Jin, Furu Wei, and Houfeng Wang. 2026. Two pathways to truthfulness: On the intrinsic encoding of llm hallucinations. *Preprint*, arXiv:2601.07422.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. 2024. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *The Twelfth International Conference on Learning Representations*.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, volume 35 of *NIPS '22*, pages 17359–17372, Red Hook, NY, USA. Curran Associates Inc.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017*.

Meta Llama Team. 2026. meta-llama/llama-3.2-3b. Hugging Face model card. Governed by the Llama 3.2 Community License and Acceptable Use Policy; accessed 2026-05-26.

Microsoft. 2026. microsoft/phi-3-mini-4k-instruct. Hugging Face model card. License listed as MIT; model card includes intended-use guidance; accessed 2026-05-26.

Buddhika Nettasinghe and Geethu Joseph. 2026. As language models scale, low-order linear depth dynamics emerge. *Preprint*, arXiv:2603.12541.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39643–39666. PMLR.

Angelos Poulis, Mark Crovella, and Evimaria Terzi. 2026. Testing the limits of truth directions in llms. *Preprint*, arXiv:2604.03754.

Qwen Team. 2026. Qwen/qwen3.5-4b. Hugging Face model card. License listed as Apache-2.0; accessed 2026-05-26.

Selvan Sunitha Ravi, Bartosz Mielczarek, Anand Kannappan, Douwe Kiela, and Rebecca Qian. 2024. Lynx: An open source hallucination evaluation model. *Preprint*, arXiv:2407.08488.

Anton Razzhigaev, Matvey Mikhalchuk, Elizaveta Goncharova, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. 2024. The shape of learning: Anisotropy and intrinsic dimensions in transformer-based models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 868–874.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522.

Or Shafran, Shaked Ronen, Omri Fahn, Shauli Ravfogel, Atticus Geiger, and Mor Geva. 2026. From directions to regions: Decomposing activations in language models via local geometry. *Preprint*, arXiv:2602.02464.

Chung-En Sun, Ge Yan, Zimo Wang, and Tsui-Wei Weng. 2026. Steer2edit: From activation steering to component-level editing. *Preprint*, arXiv:2602.09870.

Andrzej Szablewski and Marek Masiak. 2025. Activation transport operators. In *NeurIPS 2025 Workshop on Mechanistic Interpretability*. Spotlight; arXiv:2508.17540.

Magamed Taimeskhanov, Samuel Vaiter, and Damien Garreau. 2026. Towards understanding steering strength. *Preprint*, arXiv:2602.02712.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Sohan Venkatesh and Ashish Mahendran Kurapath. 2026. On the non-identifiability of steering vectors in large language models. *Preprint*, arXiv:2602.06801.

Yuanyi Wang, Yifan Yang, Su Lu, Yanggan Gu, Pengkai Wang, Wenjun Wang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, Jialun Cao, Shing-Chi Cheung, and Hongxia Yang. 2026. Geometry conflict: Explaining and controlling forgetting in LLM continual post-training. *Preprint*, arXiv:2605.09608.

Zhuofan Josh Ying, Shauli Ravfogel, Nikolaus Kriegeskorte, and Peter Hase. 2026. The truthfulness spectrum hypothesis. *Preprint*, arXiv:2602.20273.

Stanley Yu, Vaidehi Bulusu, Clayton Lau, Oscar S. Yasunaga, Cole Blondin, Vasu Sharma, Kevin Zhu, and Sean O'Brien. 2025. From directions to cones:

Multidimensional representations of propositional facts in LLMs. ACL-SRW 2025 poster. *Preprint*, arXiv:2505.21800. Non-archival poster.

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, and 2 others. 2023. Representation engineering: A top-down approach to AI transparency. *Preprint*, arXiv:2310.01405.

A Supplementary Guide and Boundaries

This supplementary appendix collects the extended controls, alternative cuts, and technical derivations that support the same three claims as the main text. It begins with a compact appendix-level synthesis, then retains the longer technical appendix for readers who want the broader audit record in one file. Labels in the retained technical appendix are namespaced only to avoid collisions with the main text.

A.1 Supplementary Scope and Reading Guide

The main text is compressed to meet the eight-page constraint, but the broader audit record remains scientifically relevant. The role of this appendix is to preserve the controls, alternate cuts, exhaustive ledgers, and auxiliary comparisons needed to evaluate the three paper-level claims without forcing them into the compact main narrative. All rows below should be read under the same discipline as the main text: positive rows matter only together with the corresponding negative, null, or leakage control.

A.2 Reproducibility and Uncertainty Summary

Table 5 collects the counts and uncertainty summaries most relevant to the main text. The table is generated from reported aggregate artifacts rather than from a new model run. Intervals are descriptive percentiles across reported aggregate rows unless the source artifact already reports a bootstrap interval; they should therefore be read as audit transparency, not as a new population-level guarantee.

The corresponding selection rules are fixed at the family level. Depth claims are selected by the full-layer held-out fidelity sweep and then stress-tested by split, occupancy, partition, basis, and branch controls. Semantic claims are selected only after the negative hidden-versus-token boundary is established; local-policy rows are reported as bounded constructive checks and are not used to override the continuation boundary. Factuality claims are organized by access family: raw-state linear, selected hidden-state, probe-rich composite, score-free geometry, template-transfer, and held-out-model families are kept separate rather than pooled into one detector claim.

A.3 Artifact Terms and Intended Use

This paper uses publicly released model checkpoints, public research benchmarks, and standard corpus slices only within the public licenses, model-card conditions, and evaluation terms provided by the original artifact sources. The submission reports aggregate metrics, fitted operators, and derived analysis artifacts; it does not redistribute third-party model weights, benchmark datasets, or raw hidden-state dumps from those third-party artifacts. Any accompanying scripts or derived research artifacts are intended only for research auditing and evaluation, and require users to obtain third-party checkpoints and datasets directly from the original providers under the original terms.

Evidence block	What it contains	Claim role
Prior-work boundary	Closest comparator families and the intended novelty boundary.	Prevents the paper from overclaiming method novelty.
Protocol boundary	CRJ fitting, exact-JVP checks, and whole-block reconstruction protocols.	Clarifies which operator is being estimated in each section.
Depth controls	Basis, split, occupancy, partition, branch, and exhaustive layer ledgers.	Supports the bounded corridor-versus-collapse reading.
Semantic controls	Cluster examples, hidden-vs-token gaps, locality, and local-policy comparisons.	Supports the descriptive-to-causal boundary.
Factuality controls	Feature-family ladders, leakage checks, transfer controls, and template sensitivity.	Supports the access-family reading and blocks detector overclaims.

Table 4: Supplementary evidence inventory. The main paper states the three core claims compactly; this appendix preserves the broader audit record needed to evaluate them.

Audit family	Rows / seeds	Main metric	Descriptive interval	Paper-facing interpretation
Semantic continuation audit	350 rows, 5 parsed seeds, 5 lexical families, 7 policies	continuation rate mean 0.0267; target rollout-positive mean 0.0022	continuation percentile range 0.0000–0.3333; maximum target rollout-positive rate 0.0833	Nonzero continuation rows exist but are sparse and weak; the semantic result remains next-token-local and bounded.
Exact-JVP naming boundary	9 anchor rows, 3 seeds per row, best partition selected per anchor	corridor exact-JVP cosine 0.5642/0.4707/0.3168 for Qwen/Llama/Phi	Q/L/P seed sd 0.0043/0.0040/0.0026; shuffled-null margins 0.5261/0.4587/0.3157	Supports the surrogate naming boundary without promoting exact pointwise Jacobian recovery.
Prompt-template factuality access	160 aggregate rows, 2 models, 4 benchmarks, 5 targets	mean cross-template AUC drop −0.0877	rowwise percentile range −0.2366–0.0190; worst-drop mean −0.2081	In-template access is real, but template transfer remains materially degraded.
Secondary template-transfer audit	32 aggregate rows, 2 models, 4 benchmarks, 2 targets	mean cross-template AUC drop −0.1051	rowwise percentile range −0.2290–−0.0231	Confirms that prompt-template sensitivity is not an isolated artifact of one target construction.
Stable-target rescue	128 non-local rescue rows	mean rescue delta +0.0250; positive rows 94/128	rowwise percentile range −0.1268–0.1444	Stable target construction mitigates some drops but does not recover template-global factuality access.
Nonlinear frozen-template baseline	6240 transfer rows, 16 model–benchmark groups, 3 seeds, 6 classifier families	source-template validation AUC mean 0.7631; cross-template AUC mean 0.6188	mean drop −0.1444; worst row drop −0.7905; nonlinear selection rate 0.6958	Nonlinear classifier capacity improves source-template selection but still does not recover template-global access.
Gate/rank bootstrap	6 model/statistic rows	partial correlation mean −0.2690	rowwise percentile range −0.7905–0.4244	Gate/rank statistics are structural markers with family sensitivity, not a standalone causal proof.

Table 5: Reproducibility and uncertainty summary for the reported multiseed and grouped-audit evidence.

Artifact	Source / citation	License / terms boundary	This paper's use
Qwen-3.5-4B	Official model card (Qwen Team, 2026).	Apache-2.0 according to the official public model card.	White-box hidden-state extraction, operator fitting, and aggregate evaluation only; no redistribution of third-party weights.
Llama-3.2-3B	Official model card (Meta Llama Team, 2026).	Llama 3.2 Community License and Acceptable Use Policy according to the official public model card.	White-box hidden-state extraction, operator fitting, and aggregate evaluation only; no redistribution of third-party weights.
Phi-3-mini-4k-instruct	Official model card (Microsoft, 2026).	MIT license according to the official public model card, together with the model-card intended-use guidance.	White-box hidden-state extraction, operator fitting, and aggregate evaluation only; no redistribution of third-party weights.
WikiText-2	Standard corpus source (Merity et al., 2017).	Use remains subject to the original corpus distribution terms.	Hidden-state collection and operator fitting only; the paper reports derived metrics rather than corpus redistribution.
HaluBench	Primary benchmark source (Ravi et al., 2024).	Use remains subject to the original benchmark access and evaluation terms.	Aggregate evaluation only; no benchmark redistribution in the submission package.
TruthfulQA	Primary benchmark source (Lin et al., 2022).	Use remains subject to the original benchmark access and evaluation terms.	Aggregate evaluation only; no benchmark redistribution in the submission package.
BoolQ	Primary benchmark source (Clark et al., 2019).	Use remains subject to the original benchmark access and evaluation terms.	Aggregate evaluation only; no benchmark redistribution in the submission package.
Derived paper artifacts	Fitted CRJ operators, aggregate result tables, summary CSVs, figures, and audit scripts produced from the above sources.	New derived artifacts from this project must still respect the access conditions of any third-party sources they depend on.	Intended for transparent research auditing; any release excludes third-party weights, benchmark contents, and raw hidden-state dumps.

Table 6: Artifact terms and intended use summary for the ARR submission.

Operation	Claim state	Supported reading	Stronger reading not supported by this block	Key numbers
Description	supported	Middle-layer corridor and late collapse are real operator-surrogate regimes in the tested families.	Exact pointwise Jacobian recovery or a universal best-layer rule.	Exact-JVP means: Qwen L7 0.5642, Llama L14 0.4707, Phi L11 0.3168; learned-feature partition deltas stay at $\sim 10^{-3}$ scale.
Control	bounded	Hidden-response local structure can be trust-region faithful and sometimes policy-conditioned.	Clean global concept dials, unique semantic directions, or robust continuation-level steering.	945 contrasts: 448 decisive negatives, 292 mixed, 184 phrase-level support, 21 next-token only; low-alignment directions recover at least 80% of the target V1 effect in 47.6% of groups.
Transfer	closed boundary	Selected and probe-rich access families carry real signal, but transfer remains template-sensitive and family-sensitive.	Template-global truth geometry, score-free truth axis, or detector-leaderboard claims.	Mean cross-template drops remain materially negative; stable targets improve 73.4% of non-local rows, but mean rescue is only about +0.0250 AUC-drop points.

Table 7: Operation-level appendix summary synthesized from the reported audit evidence.

A.4 Operation-Level Appendix Summary

The main text already separates readout, control, and transfer at the level of supported readings. Tables 7 and 8 make that role explicit in the appendix by compressing the reported evidence state into one operation-level summary and one condition-level matrix. These tables are synthesis devices only: they do not add new evidence beyond the reported audits already cited elsewhere in the appendix.

A.5 Prior-Work Boundary

The novelty claim is not that local linearization, Jacobian analysis, steering, or truth probing are new. The claim is that a single population-level, cluster-representative hidden-to-hidden operator stack is used to audit where linear access remains descriptive, where it fails causally, and where it fails to transfer. Table 9 records the closest comparison families.

A.6 CRJ Protocols and Method Boundary

A cluster-representative Jacobian (CRJ) is the region-conditioned best linear predictor,

$$J_k^\star = \arg\min_J \mathcal{L}_k(J),$$
$$\mathcal{L}_k(J) = \mathbb{E}\left[\|Y - \bar{y}_k - J(X - \bar{x}_k)\|_2^2 \mid a(X) = k\right].$$

It is not an exact pointwise derivative. For attention-containing blocks, sequence-context marginalization remains even as the token-space region shrinks. For position-wise submodules such as the MLP/FFN, exact autograd JVP comparisons are meaningful as a sanity check. Those exact-JVP rows are a validation target for the CRJ naming boundary, not the main estimand used in the depth, semantic, or factuality audits. Table 10 separates the three operator protocols used in the project.

The factuality family names are also audit shorthands rather than theoretical primitives. For exam-

Operation	Condition	Status	Supported reading	Evidence block
Description	exact-JVP naming boundary	supported	CRJ remains a population-level surrogate; calibration succeeds best in corridor anchors and weakens early/late.	isolated-MLP exact-JVP calibration
Description	learned-feature partition rescue	closed boundary	Off-hook SAE and exact-hook transcoder partitions remain inside the raw/single/random envelope.	learned-feature and hook-alignment partition audits
Description	late-collapse mechanism intervention	bounded	Corridor fragility to dominant-PC sharpening is strong; simple rescue manipulations remain unsupported, including the corrected three-seed PC/rebalance repeat.	dominant-PC perturbation audits
Control	target-control prospective suite	bounded	Constructive evidence survives mainly as phrase-level or next-token-local pockets rather than robust control.	prospective target-control benchmark
Control	continuation-level continuation audit	closed boundary	Llama retains sparse weak continuation pockets; Qwen does not; the result stays next-token-local.	multiseed continuation audit
Control	direction uniqueness	closed boundary	Token-facing effect magnitude is not uniquely identified by the target V1 direction.	direction-equivalence audit
Transfer	in-template access	supported	Hidden-state access remains real within fixed prompt templates.	fixed-template hidden-state access audit
Transfer	cross-template transfer	closed boundary	Cross-template transfer degrades materially enough to block template-global truth-geometry wording.	cross-template transfer audit
Transfer	stable-target reconstruction	bounded	Stable targets explain part of the prompt sensitivity, but recovery is partial and slice-specific.	stable-target reconstruction audit
Transfer	score-free geometry	closed boundary	Score-free geometry remains weak even when richer probe-rich families transfer better.	grouped geometry-versus-probe audit

Table 8: Condition-level appendix summary for the reported evidence state.

ple, `core_minus_direct` removes only the direct scalar score columns `s0_score`, `s1_score`, and `s2_score`; it does not remove every probe-derived summary. This distinction matters for interpreting detector transfer.

B Evidence Extensions

B.1 Depth Evidence

B.2 Anchor Layers and Basis Sensitivity

Across Qwen-3.5-4B, Llama-3.2-3B, and Phi-3-mini, the main depth sweep reports a middle-layer corridor followed by late collapse. The peak bands are Qwen layers 5–19 with peak cosine 0.56 at layer 7, Llama layers 4–15 with peak 0.50 at layer 14, and Phi-3 layers 8–15 with peak 0.27 at layer 11. The absolute levels and peak locations differ, so the claim is a regime pattern in the tested audit family rather than a universal best layer.

On Qwen layer 12, a K=64 diagonal operator achieves cosine fidelity 0.473, close to a single global operator at 0.487. This means the corridor can be captured by simple local scaling in that setting; it does not mean the whole network is globally axis-separable. At Qwen layer 31, diagonal and axis-relabeling controls remain near collapse, while an SVD basis partially recovers fidelity.

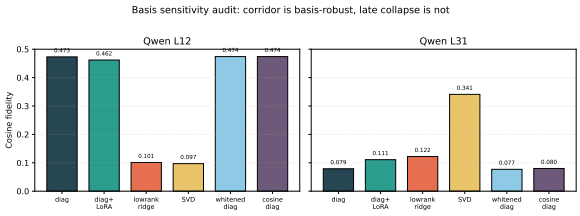

Figure 2: Basis-sensitivity audit on Qwen-3.5-4B. The favorable middle-layer corridor stays in nearly the same fidelity band under diagonal, cosine, and whitened relabelings, while late-layer structure remains fragile: diagonal and simple relabelings stay near collapse, LoRA and low-rank ridge recover only modestly, and an SVD basis gives the strongest late-layer rescue without restoring a general late-law.

B.3 Residual Mismatch and Centroid Distance

This appendix checks whether late collapse is merely a clustering artifact. For Qwen, band-averaged relative mismatch energy is 0.951 in early layers, 0.766 in the sweet spot, and 1.004 in the late band. For Llama the corresponding values are

Prior family	What it establishes	Boundary for this paper	Role in the ARR framing
Relation decoding / linear relational embeddings (Hernandez et al., 2024)	Some factual computations admit first-order subject-to-object approximations, but not all relations do.	We audit where operator fidelity survives across layers, regions, and downstream access families rather than only for a fixed relation family.	Closest factual-computation comparator.
Prompt-conditioned Jacobians and transport maps (Golden, 2025; Nettasinghe and Joseph, 2026; Szablewski and Masiak, 2025)	Fixed-context, low-order, or residual-transport operators can explain or control aspects of LLM computation.	We fit population-level cluster-representative hidden-to-hidden operators and audit held-out fidelity plus downstream failures.	Closest method comparator.
SAE and local-region decompositions (Huben et al., 2024; Bricken et al., 2023; Farnik et al., 2025; Shafran et al., 2026)	Activation features, sparse computation bases, and local regions expose interpretable structure.	Our primary object is an input-output operator with fidelity, residual, gate, and token-effect metadata, not a feature dictionary.	Motivates piecewise structure.
Depth geometry and anisotropy (Razzhigaev et al., 2024; Godey et al., 2024; Haim and McNamee, 2026)	Layer depth reorganizes representation geometry and prediction structure.	We measure dynamic local operator fidelity and add split, occupancy, basis, and branch controls.	Structural backdrop.
Activation steering and patching (Turner et al., 2023; Rimsky et al., 2024; Taimeskhanov et al., 2026; Venkatesh and Kurapath, 2026; Li et al., 2026)	Linear interventions can work but are sensitive, non-identifiable, and sometimes off-target.	We decompose hidden-response, tracked log-probability, boundary-token, and local-policy effects.	Defines causal-control target.
Truth probes and hallucination detectors (Marks and Tegmark, 2024; Liu et al., 2024; Poulis et al., 2026; Luo et al., 2026; Ying et al., 2026)	Truthfulness can be encoded and detected under supervised, layer-, pathway-, or dataset-specific settings.	We separate raw-state linear, selected hidden-state, probe-rich, and score-free geometry access families.	Defines factuality boundary.

Table 9: Closest prior-work families and the intended novelty boundary. This table is retained because it prevents the ARR version from overclaiming method novelty.

0.935, 0.778, and 1.005. Cosine fidelity and residual mismatch are almost perfectly anti-correlated (Qwen $r = -0.980$, Llama $r = -0.984$), while residual mismatch and centroid distance are weakly correlated (Qwen $r = 0.260$, Llama $r = 0.195$). Phi-3 has lower absolute fidelity but the same qualitative pattern: band cosine values 0.059, 0.207, 0.223, and 0.026, with cosine versus residual correlation $r = -0.992$. These values support the statement that centroid distance alone is a poor explanation of late collapse.

B.4 Branch Controls and Failed Simple Rescues

Branch-decomposition controls sharpen the depth result. Qwen layer 12 is MLP/block-led with cosine 0.237/0.230 against attention 0.013, while by layer 31 all branches are near collapse (0.036, 0.043, 0.048). Llama layer 12 is also MLP-led (0.325 versus attention 0.029), but by layer 27 attention overtakes MLP (0.157 versus 0.053). Pythia mirrors the MLP-led pattern, whereas OPT provides an attention-led middle-depth counterexample.

Light branch-scaling canaries move fidelity in the predicted branch-specific directions for some families. Llama layer 12 improves by about +0.096 under MLP amplification; Llama layer 27 improves under MLP attenuation; OPT layer 14 improves modestly under attention amplification. A raw-scaling pass extends the sign-flip pattern to Pythia (+0.040 at layer 14 under MLP amplification and +0.015 at layer 30 under MLP attenuation), while Qwen remains comparatively stable. Thus the branch account is family-dependent.

The causal-identification sequel is deliberately bounded. Norm-preserved sharpening of the primary PC1 direction reliably damages corridor fidelity across Qwen, Llama, and Phi-3, with mean Δcosine -0.0434 at $\epsilon = 0.02$ and -0.0190 at $\epsilon = 0.05$. The inverse manipulations do not reverse late collapse: primary-PC1 flattening in late rows remains negative (-0.0408 and -0.0306), trace re-

Protocol	Object fitted or computed	Scope
CRJ fitting	Population-level conditional best linear predictor, usually LoRA-factorized or diagonal.	Main depth, semantic, and factuality audits.
Branch-local exact JVP	Pointwise autograd Jacobian-vector products per MLP submodule.	Exact-JVP boundary checks only.
Whole-block reconstruction	Block-level input-output surrogate with shared clustering and per-branch fits.	Branch-decomposition controls.

Table 10: Linear-operator protocols. The term ''Jacobian'' is operational in the CRJ rows and literal only in exact-JVP submodule checks.

Family	Corridor band	Late diagonal	Best late recovery
Qwen-3.5-4B	0.473–0.476	0.067–0.079	SVD 0.341
Phi-3-mini	0.248–0.253	0.013	LoRA/ridge 0.068
Llama-3.2-3B	0.474–0.477	0.087–0.126	SVD 0.313

Table 11: Cross-family basis-sensitivity anchors. The corridor is stable across simple basis relabelings, while late-layer structure is basis-sensitive and the strongest recovery varies by model.

balancing is negative (-0.0219 and -0.0163), and flattening plus secondary rebalancing is worse (-0.0618 and -0.0534). The supported mechanism wording is therefore an identified fragility plus a failed simple-reversal family, not a single-cause theorem.

B.5 Occupancy and Split Hygiene

Late collapse is not explained by under-occupied local fits in the reported sweeps. Median train occupancy changes only from 14.0 to 14.5 in Qwen, 14.0 to 13.5 in Llama, and 8.0 to 8.5 in Phi-3 when moving from the middle corridor to the late band. Layer-wise train-count correlations are non-negative. A separate occupancy-controlled K-sweep shows that once occupancy is held near 44–50 samples per cluster, increasing local resolution improves held-out MSE rather than destabilizing the fit.

B.6 Additional Control Audits

An additional audit block targeted three reviewer-shaped gaps that remained scientifically relevant: (i) semantic-label and attention–MLP-concat partition families, (ii) a low-rank context-bridge audit, and (iii) a cleaner retrospective semantic predictor summary. The role of these additions was to

Regime	Official cos.	Within-train cos.	Mean rel. L2
Pre-corridor	0.1650	0.1654	0.9783
Corridor	0.3330	0.3339	0.9104
Late collapse	0.0648	0.0681	0.9960

Table 12: Split-hygiene summary over 18 arms. The corridor ordering survives both official and within-train protocols; Phi-3 remains a lower-fidelity cautionary family.

test whether the remaining loopholes could materially overturn the paper's bounded validity-envelope reading.

Two implementation notes matter for interpretation. First, the executable Qwen arm in this partition stress test used 'Qwen/Qwen2.5-3B' rather than 'Qwen/Qwen3.5-4B', so this row should be read as a same-family stress test rather than as a direct replication of the original Qwen-3.5 line. Second, the low-rank bridge rows are reported from the corrected implementation used for this appendix.

B.7 Semantic Evidence

B.8 Descriptive Cluster Structure

The semantic section begins with a genuine descriptive success. At K=2000 on 100,000 WikiText-2 tokens, Qwen has 1997 non-empty clusters and Llama has 1980. Representative clusters include Llama cluster 1662 with 2138 samples for surname prefixes such as *Pe*, *Van*, *Sch*, and *Mc*; Qwen cluster 1327 with 964 samples for function words such as *the*, *and*, and *in*; Qwen cluster 1717 for date and year fragments; and Qwen cluster 1523 for appearance and career verbs such as *appeared* or *starred*.

The singular spectra are moderately anisotropic rather than rank-1. The median σ_1/σ_2 ratio is about 1.16; 32.4% of clusters exceed 2 and 11.3% exceed 5. Extreme ratios above 10 are disproportionately concentrated in low-occupancy clusters with fewer than 10 samples, while clusters with at least 20 samples usually remain multi-dimensional.

B.9 Boundary Suppression and Descriptive-Causal Gap

The representative Qwen layer 12 cluster 27 intervention shows why semantic purity is not enough. The first singular direction appears family/relational by top-token inspection, but intervention changes are dominated by boundary-token suppression.

Other cluster examples prevent this from becoming a single-hand-picked example. Qwen cluster 381, selected as a pure family candidate, produces

Audit target	New audit arms	Quantitative result	Supported reading
Partition stress test	'supervised_semantic_lexical', 'attention–MLP-concat', plus matched controls on Qwen2.5-3B and Llama-3.2-3B at pre/corridor/late anchors	'attention–MLP-concat' is best in 3/12 model-layer-split cells, with mean Δcosine about −0.002 versus the best alternative partition; 'supervised_semantic_lexical' is best in 1/12 cells, with mean Δcosine about −0.0086.	Richer region definitions can be competitive in individual cells, but they do not create a new privileged semantic partition regime.
Low-rank bridge stress test	diagonal, diagonal+LoRA, and centered-IO low-rank fits over whole-block / attention-branch and same-token / cross-token rows	Best-of-family exact-JVP remains high for whole-block same-token (∼ 0.75 corridor, ∼ 0.84 late) but low for whole-block cross-token (∼ 0.03 corridor, ∼ 0.11 late) and weak for attention-branch rows (∼ 0.06–0.14). Mean exact-JVP gain of the best low-rank fit over 'jvp_diag' is only about +0.004.	Low-rank capacity sharpens some centered-IO and secant summaries, but does not rescue the weak cross-token or attention-branch exact-JVP boundary.
Semantic retrospective summary	pooled intervention rows from the semantic audit bank	Strongest pooled descriptive correlation with control-adjusted magnitude is about $\rho = 0.62$; strongest same-scope correlation with specificity is only about $\rho = 0.21$.	Descriptive quality predicts whether a direction moves the model much more than whether it preserves semantic specificity over controls.

Table 13: Additional control summary. These audits close remaining interpretive loopholes without changing the main-text regime map.

Target family	Mass change	Interpretation
Family/relational	$+1.39 \times 10^{-5}$	Negligible activation
Boundary/punctuation	-5.70×10^{-3}	Strong suppression

Table 14: Qwen layer 12 cluster 27 V1 intervention. The dominant token-facing effect is boundary suppression, about 400 times larger in raw mass than the positive-family increase.

Response family	Rows	Mean cos.	Mean rel. L2
Hidden delta basis	144	0.9892	0.1645
Tracked log-prob delta	144	0.5586	1.0796

Table 15: Local parameter-edit linearization audit. Hidden-response deltas are highly linear, but tracked log-probability responses are much weaker.

semantic drift toward tokens such as *released*, *fired*, and *parted ways*; Llama cluster 1549 is a matched family-like comparison around *mother* and *wife*. Across 132 non-random audited directions, descriptive quality tracks control-adjusted effect magnitude moderately well ($\rho \approx 0.62$ for the strongest NLL-based measure), but remains a weak predictor of specificity over matched controls ($\rho \approx 0.15$ to 0.26).

The model breakdown in the tracked log-probability rows is also heterogeneous: Qwen averages 0.7432, Llama 0.5687, and Phi-3 0.3640, while hidden-response cosine remains near 0.99 in all three. This is the access-family gap stated in the main paper.

B.10 Atlas Locality and Token-Family Mixedness

The partially recovered edit-atlas pass supports locality without completing a stronger two-tangent mechanism claim. Table 16 summarizes the reported same-cluster versus cross-cluster transfer rows.

The token-family analysis is mixed rather than cleanly semantic. Positive-family absolute changes are only modestly larger than boundary-token changes on average: mean positive/boundary absolute ratio 1.3151 and median 1.1813. The fraction of examples where positive absolute change exceeds boundary absolute change is 0.5515. Sign-level fractions remain close to balanced: the positive-family mean is greater than zero in 0.5087 of rows, and the boundary-family mean is less than zero in 0.4855.

Model	Response	Same cos.	Cross cos.	Gap	Same rel. L2
Llama	Hidden	0.8524	0.1515	0.7008	0.6158
Llama	Tracked	0.4235	0.0809	0.3427	1.4778
Phi-3	Hidden	0.8995	0.1353	0.7642	0.4375
Phi-3	Tracked	0.3336	0.0377	0.2959	1.5217
Qwen	Hidden	0.8243	0.2793	0.5451	0.6869
Qwen	Tracked	0.4996	0.1034	0.3962	1.3694

Table 16: Recovered edit-atlas locality. Same-cluster hidden-response transfer is much stronger than cross-cluster transfer, but tracked token-facing transfer remains weaker and higher-error.

B.11 Constructive Pockets and Prospective Negatives

Local conditional policies give a bounded constructive signal. On Qwen layer 12 cluster 27, the target-versus-control boundary-gap magnitude improves from about 0.00078 under the global target policy to 0.00849 under a conditional target-only policy. A routed-cluster policy improves contrast-level specificity from 0.0024 to 0.0197 in the initial comparison. For Qwen clusters 1814 and 1592, the best local policy is routed-cluster V1, with specificity-positive gaps 0.001058 and 0.000710 and specificity-boundary gaps 0.000300 and 0.000381. A matched Llama comparison remains much weaker at the 10^{-5} to 10^{-4} scale.

The fixed-selector family sweep broadens this only modestly. Across five lexical families at the same Qwen layer, local policies retain positive next-token specificity in four families and beat the best global policy in three, but sentiment does not replicate and all continuation evaluations remain soft-only. In the cross-layer prospective benchmark, the corridor has the strongest absolute local lexical effect on average (about 3.22×10^{-4} versus 6.99×10^{-5} pre-corridor and 2.88×10^{-5} late), but pre-corridor layers beat their own matched global baselines more consistently. Frozen-casebook rollouts keep hard continuation target and control hits at zero; only weak soft target-mass shifts remain, such as Qwen cluster 1592 at $+2.1 \times 10^{-5}$ and cluster 1814 at $+3.1 \times 10^{-5}$.

The larger prospective semantic suite makes the negative boundary explicit. Across 945 target-control contrasts, the outcome labels are 448 decisive negatives, 292 mixed cases, 184 phrase-level support cases, and 21 next-token-only positives. Phrase-level support is concentrated mostly in Qwen occupation and temporal cases under corridor or selected late-layer settings. Many pre-corridor and late-collapse settings remain negative, conditional centroid pulls can decisively damage negation, kinship, and sentiment targets, and rollout selectivity is essentially absent in the top rows.

The contrastive mean-difference baseline reinforces the same caution. On the cluster-27 comparison, the CAA-style mean-difference vector is moderately anti-aligned with the target CRJ direction (cosine -0.446). At $\alpha = 1.0$, target V1 yields a positive contrast delta of +0.001188 while the contrastive baseline yields -0.000420; at $\alpha = 2.0$, the corresponding values are +0.004197 and -0.000124. The baseline is therefore not a substitute for the CRJ direction in this comparison, although both remain local, soft-only, and top-1-stable.

B.12 Factuality Evidence

B.13 Raw-State Linear Baselines

The factuality section intentionally starts from a weak baseline. On HaluBench, the tested raw-state linear probes are near chance: Qwen best layer 0 AUC 0.466 with balanced accuracy 0.492 and FAIL rate 0.086; Llama layer 27 AUC 0.514 with balanced accuracy 0.483 and FAIL rate 0.739; Phi-3 layer 0 AUC 0.474 with balanced accuracy 0.489 and FAIL rate 0.122. This is the raw linear-access boundary, not a claim that no internal factuality signal exists.

As calibration, SelfCheckGPT-style baselines on the same 1200-example HaluBench slice reach AUC 0.606 for `selfcheck_nli`, 0.451 for `selfcheck_bert`, and 0.503 for `inside_eigenscore`. These numbers are included to orient the benchmark, not to turn the paper into a hallucination-detection leaderboard.

B.14 Selected Hidden-State and Probe-Rich Access

Selected hidden-state probes recover benchmark-sensitive signal under grouped source-, pair-, and length-style holdouts. Representative rows are Qwen HaluBench 0.775/0.712, Llama HaluBench 0.727/0.660, Qwen TruthfulQA 0.730/0.673,

Feature family	Contents	Interpretation
`core_probe`	Layerwise probe trajectories, norms, `prob_fail`, `correct`, layer indices, and selected-layer counts; excludes `traj_*`.	Strong supervised probe-rich access.
`core_minus_score`	`core_probe` after removing score columns.	Tests dependence on explicit scores.
`core_minus_direct`	Removes only `s0_score`, `s1_score`, and `s2_score`; retains other probe summaries.	A restricted direct-score ablation, not removal of all probe-derived data.
`geometry_only`	Mismatch, scale, and gate summaries.	Score-free geometry access.
`trajectory_only`	`traj_*` summaries only.	Pure trajectory summary access.
`core_plus_geometry`	Probe-rich core plus geometry.	Tests incremental geometry value.
`metadata_only`	Non-model metadata fields.	Leakage/control family.

Table 17: Factuality feature-family definitions. This table is essential because the strongest detector rows are probe-rich and should not be described as pure geometry.

Llama TruthfulQA 0.772/0.687, Qwen BoolQ 0.628/0.626, and Llama BoolQ 0.634/0.628, where the two values summarize grouped variants. Grouped maxima include Qwen HaluBench at 0.826 pair-grouped and Llama TruthfulQA at 0.786 pair-grouped. The HaluEval-QA anchor is stronger but easier, so it remains a secondary appendix result rather than a headline claim.

B.15 Leakage, Score-Free, and Transfer Controls

The detector-ladder result is a feature-family split. Metadata-free `core_probe_nometa` reaches 0.973 in pooled-model grouped evaluation and 1.000 in pooled-source grouped evaluation, while `core_minus_direct_nometa` falls near chance at 0.499 and 0.515. In case-level row-id grouping, `core_probe_nometa` reaches 1.000 RF/HGB for factuality and 0.991 RF for probe-failure, while geometry and trajectory families remain around 0.506. Grouped geometry-delta AUC is uniformly negative in the reported audit (-0.2845 to -0.4499), the best score-free pooled grouped result is about 0.505, and the best within-model grouped geometry row is about 0.530.

The held-out-model audit is the strongest transfer guardrail. For factuality, Qwen and Phi-3 reach 1.000 with probe-rich subfamilies including `core_probe`, `core_minus_direct`, `core_minus_score`, `core_plus_geometry`, and `core_plus_geometry_plus_trajectory`, while geometry and trajectory stay at 0.500. Held-out Llama is weaker: `core_probe` reaches 0.7708 and `core_plus_geometry` reaches 0.8584. Probe-failure rows are also family-sensitive, with Qwen probe-rich rows around 0.6375 to 0.746, Phi-3 around 0.818 to 0.892, and Llama around 0.829; geometry and trajectory are weaker but not uniformly zero in that auxiliary target.

A prompt-template audit tightens that transfer boundary further. Within a fixed prompt template, hidden-state access remains real, but cross-template transfer degrades materially: mean AUC drops range from about -0.02 to -0.23, and the worst reported drop is about -0.42. The target itself is not template-invariant either. For `model_incorrect`, the weakest template-to-template agreement in the reported audit is 0.110 on Llama TruthfulQA. This closes off the tempting overread that the factuality signal here is a template-global truth geometry. The supported reading is narrower: benchmark-, family-, and prompt-template-sensitive access families.

The stable-target reconstruction audit shows that this prompt sensitivity is partly target-definition instability, but not only that. Rebuilding `model_incorrect` into majority-, supermajority-, and unanimous-template targets improves 73.4% of non-local rescue rows relative to the local `model_incorrect` baseline, with mean rescue delta about +0.0250 and best rescue about +0.1444 on Llama HaluEvalQA. At the same time the rescue remains mixed across benchmarks: benchmark means are still slightly negative on Llama HaluBench and Qwen BoolQ, and even improved slices retain nontrivial cross-template AUC loss. The correct conclusion is therefore bounded. Stable targets remove one real confound, but they do not recover a template-global hidden-state factuality rule.

Within-model nonlinear detectors are high, but raw-coordinate leave-model-out transfer collapses. The within-model factuality rows are Qwen LogReg 0.885, MLP 0.998, RF 1.000, HGB 1.000; Llama 0.916, 1.000, 1.000, 1.000; and Phi-3 0.884, 0.991,

Audit	Key numbers	Consequence
Raw-state linear HaluBench	Qwen 0.466, Llama 0.514, Phi-3 0.474 AUC.	Tested raw-state linear family is too weak.
Selected hidden-state extension	Representative grouped rows: Qwen HaluBench 0.775/0.712, Llama TruthfulQA 0.772/0.687, BoolQ about 0.63.	Hidden-state access can recover benchmark-sensitive signal.
Feature-leakage ablation	`core_minus_direct` and geometry-only / trajectory-only drop near chance in grouped settings.	High detector rows are probe-rich, not pure geometry.
Score-free geometry	Pooled grouped best about 0.505; best within-model grouped about 0.530.	Score-free geometry remains weak.
Held-out-model metadata-free	Probe-rich subfamilies reach up to 1.000 for held-out Qwen/Phi-3 and 0.77 to 0.86 for held-out Llama; geometry-only / trajectory-only stay at chance.	Transfer is family-sensitive and probe-rich.
Permutation sanity	Grouped hidden rows collapse to about 0.51 and grouped detector-core rows to about 0.53 under label permutation.	Rules out a simple split-implementation leak.

Table 18: Factuality audit inventory. The appendix preserves the guardrails against reading the detector as a pure truth-geometry result.

Object for K = 40k, N = 500k, d = 3072, r = 16	Naive	Flash-Jacobian
Distance matrix	40GB	Not materialized
Full Jacobians	755GB	Not stored
Low-rank operators	–	7.9GB
Data	6.1GB	6.1GB
Centroids	0.5GB	0.5GB
Approx. total	801GB	14.5GB

Table 19: Memory accounting retained from the preprint. The system contribution is an audit-enabling cost envelope, not a claim that every downstream audit is cheap.

1.000, 1.000. Raw-coordinate cross-model pooled evaluation gives LogReg 0.464, MLP 0.473, RF 0.513, and HGB 0.493. The right interpretation is therefore access-family sensitivity, not a universal truth direction.

C Technical and Scale Details

C.1 System and Complexity Details

C.2 Resource Envelope

The reported Flash-Jacobian fit-cost envelope uses a single NVIDIA L4 with 23.7GB VRAM, 53GB RAM, and 112.6GB storage under PyTorch/CUDA. The synthetic CPU checks use an Apple M1 with 16GB RAM. The main Qwen layer-12 K=2000, rank-16 fit over about 98,412 tokens completes in about 155–157 seconds on the L4 for the fitting stage, excluding upstream tokenization and activation extraction. This single-L4 row is specific to the fit-only cost envelope. Some later closure and transfer follow-up packages cited in the paper were run on other GPU hosts, including multi-GPU RunPod environments recorded in the package-local analysis notes.

C.3 Hyperparameters and Data

Unless a specific experimental block says otherwise, the default configuration uses LoRA rank $r = 16$, ridge regularization $\lambda = 10^{-2}$, maximum iterations 20, convergence tolerance 10^{-4}, FP16 LLM activations, FP32 synthetic checks, Xavier LoRA initialization, and random centroid initialization. The primary open checkpoints are Qwen-3.5-4B, Llama-3.2-3B, and Phi-3-mini-4k-instruct from their official public model-card distributions (Qwen Team, 2026; Meta Llama Team, 2026; Microsoft, 2026). The main corpus and benchmark artifacts are WikiText-2 (Merity et al., 2017), HaluBench (Ravi et al., 2024), TruthfulQA (Lin et al., 2022), and BoolQ (Clark et al., 2019). WikiText-2 collection uses 1000 passages averaging roughly 512 tokens, records hidden-state input-output pairs, and subsamples to 100k tokens for dense K=2000 runs.

C.4 Scaling, Resolution, and Ablations

The synthetic scaling validation is linear in the intended bottleneck range: the small row $N = 50$k, $K = 1000$ takes 61.3 seconds; medium $N = 200$k, $K = 4000$ takes 234.7 seconds; and large $N = 500$k, $K = 10000$ takes 591.2 seconds. The K-N sweep reports the following MSE/occupancy pairs: K=25 gives 49.20/189.03; K=50 gives 54.69/212.36; K=100 gives 68.29/142.99; K=200 gives 82.33/119.69; K=400 gives 164.81/112.56; K=800 gives 132.73/111.65; K=1600 gives 138.13/103.77. This sweep is used only to motivate the resolution-occupancy tradeoff, not as a universal optimal-K rule.

At K=2000, increasing rank improves test MSE

Config	MSE	Improvement	R^2	Time (s)
K=1, r=16	0.00272	36.5	0.934	11.1
K=250, r=16	0.00229	46.5	0.944	36.1
K=500, r=16	0.00217	49.4	0.947	49.6
K=1000, r=16	0.00215	49.9	0.948	84.0
K=2000, r=16	0.00213	50.4	0.948	154.4

Table 20: Qwen layer-12 resolution ablation. The full dense run improves over smaller K but is not the only setting that recovers the corridor.

from 0.00236 at r=1 to 0.00213 at r=16, with only small fit-time variation around 153–156 seconds. Seed stability is high: three Qwen layer-12 K=2000 runs report test MSEs 0.002131, 0.002198, and 0.002167, with mean 0.002165 and standard deviation 3.36×10^{-5}; test cosine is 0.501, 0.498, and 0.500.

C.5 Exact-JVP Boundary Checks

The exact-JVP supplement clarifies what CRJs do and do not recover. In a toy ReLU setting, a regime-exact oracle with K=2001 gives cosine 1.000, relative L2 0, norm ratio 1, and relative Frobenius error 0. Under a cap of K=201, cosine drops to 0.513 and relative L2 is 1.010; an x-KMeans plus JVP-ridge row with K=50 gives cosine 0.656 and relative L2 0.777. This validates that exact piecewise-linear recovery is possible in a controlled toy when the true regions are matched, but it does not transfer automatically to transformer blocks.

On an isolated Qwen layer-12 MLP submodule, low-rank ridge reaches cosine 0.096 and relative L2 0.997; diagonal reaches cosine 0.475 and relative L2 0.880; diagonal plus LoRA reaches cosine 0.465 and relative L2 0.884. The conclusion is narrow: CRJs are local surrogates for regime auditing, while exact differential recovery remains a separate and stricter target.

A full-depth calibration pass removes the old anchor-only caveat for the isolated-MLP surrogate-to-differential check. Under the multi-seed exact-JVP anchor calibration, the best anchor partition gives Qwen layer 7 cosine 0.564 ± 0.004 with a shuffled-null margin 0.526, Llama layer 14 cosine 0.471 ± 0.004 with margin 0.459, and Phi-3 layer 11 cosine 0.317 ± 0.003 with margin 0.316. The corridor-versus-late separation therefore survives the full-layer expansion rather than depending on a small hand-picked anchor set.

C.6 Scale Checks, SAE Comparison, and Negative Families

Family-diversity checks support the depth account non-uniformly. Mistral-7B, Qwen2.5-7B, and Gemma-2-9B reproduce mid-depth operator-fidelity peaks with late degradation; representative best partitioned cosines are 0.329, 0.449, and 0.448, while late anchors are 0.110, 0.081, and 0.019. Their best grouped factuality AUCs are 0.798, 0.844, and 0.830. Falcon-7B-Instruct is a negative family control with best partitioned depth cosine 0.254 and weak grouped factuality AUC around 0.559–0.579.

The SAE comparison remains a bounded appendix result. On the Qwen HaluBench layer-31 comparison, the best SAE-to-surrogate cosine is 0.0602 and the mean absolute cosine is 0.0213. This suggests near-orthogonality between that SAE representation and the factuality surrogate in the tested setting, but it is not a broad SAE no-go theorem.

The ARR main text deliberately removes or compresses the following material: the full prior-work boundary table; the operator-family and basis-sensitivity tables; the residual-mismatch and centroid-distance derivations; the detailed branch-scaling and failed-reversal rows; the semantic cluster examples and spectral occupancy discussion; the 132-direction descriptive-causal audit; the full 288-row parameter-edit audit; the partial edit-atlas rows; the prospective semantic-suite breakdown; the raw-state, selected-layer, feature-family, leakage, transfer, and score-free factuality ladders; the memory, scaling, K-N, rank, seed, and exact-JVP details; and the family-diversity, SAE, and Falcon cautionary checks. These elements are retained here because they support the same three main claims: depth-local operator fidelity is regime-bound; descriptive semantic directions do not imply clean global control; and factuality access depends on the feature family and transfer protocol. The remaining subsections retain the longer technical appendix material, with labels prefixed under 'full:' only to avoid collisions inside this single file.

C.7 Woodbury Identity Derivation

For a cluster with n_k samples and feature dimension d, the least-squares solution for B_k is:

$$B_k = W_k(\Delta X_k \Delta X_k^\top + \lambda I)^{-1} \Delta X_k \qquad (3)$$

when $n_k < d$, where $W_k = (A_k^\top A_k + \lambda I_r)^{-1} A_k^\top \Delta Y_k^\top \in \mathbb{R}^{r \times n_k}$. This identity allows us

to invert an $n_k \times n_k$ matrix instead of a $d \times d$ matrix, dramatically reducing computational complexity for high-dimensional models.

C.8 Flash-Jacobian Algorithm

Algorithm 1 FLASH-JACOBIAN Fitting

Require: Data $(X, Y) \in \mathbb{R}^{N\times d} \times \mathbb{R}^{N\times d}$, K, r, λ, $T_{\max}$, τ

Initialize centroids C (random), $A \sim \mathcal{N}(0, 0.01)$, $B \sim \mathcal{N}(0, 0.01)$
for $t = 1, \ldots, T_{\max}$ **do**
  $a \leftarrow \texttt{FlashAssign}(X, C)$ {No $N\times K$ matrix}
  Update $C, \bar{Y}$ {Centroid update}
  $\texttt{idx}, \texttt{bdry} \leftarrow \texttt{SortInverse}(a)$ {$O(K)$ atomics}
  for $k = 1, \ldots, K$ **do**
    Extract $\Delta X_k, \Delta Y_k$ using $\texttt{idx}, \texttt{bdry}[k]$
    Update $(A_k, B_k) \leftarrow \texttt{RidgeALS}(\Delta X_k, \Delta Y_k, r, \lambda)$ {Woodbury if $n_k \ll d$}
  end for
  Compute loss $\mathcal{L}_t$
  if $|\mathcal{L}_t - \mathcal{L}_{t-1}| < \tau$ **then**
    break
  end if
end for
return $\{(A_k, B_k, \bar{x}_k, \bar{y}_k)\}_{k=1}^{K}$

To operationalize this algorithm mathematically, for a given assigned cluster k containing exactly n_k discrete samples, we form the centered structural matrices $\Delta X_k \in \mathbb{R}^{n_k\times d}$ and $\Delta Y_k \in \mathbb{R}^{n_k\times d}$ by sequentially stacking the extracted row vectors $\Delta x_i^\top$ and $\Delta y_i^\top$. With $A_k \in \mathbb{R}^{d\times r}$ and $B_k \in \mathbb{R}^{r\times d}$, the stacked prediction is $\Delta Y_k^\top \approx A_k B_k \Delta X_k^\top$. The updates alternate between:

$$A_k \leftarrow \Delta Y_k^\top Z_k^\top (Z_k Z_k^\top + \lambda I_r)^{-1}, \quad Z_k = B_k \Delta X_k^\top \tag{4}$$

$$B_k \leftarrow W_k (\Delta X_k \Delta X_k^\top + \lambda I_{n_k})^{-1} \Delta X_k \tag{5}$$

where the B_k update uses the Woodbury identity when $n_k \ll d$.

C.9 Supporting Audits

C.10 Model Identity / Provenance Fingerprints

One downstream application explored here is model-identity and provenance auditing. Official Qwen3.5-4B-Base and its unsloth variant are strongly separable, indicating a real lineage-like packaging fingerprint; utility-preserving laundering does not erase that signal; and row-local trajectory features recover essentially the same fingerprint as endpoint summaries, showing that the signal is robust rather than path-fragile. An OOD false-positive audit further shows that the official-vs-unsloth classifier does not simply saturate on unrelated matched-pair controls. At the same time, a same-family scaling control demonstrates that prompt geometry also carries an extremely strong scale signature. Taken together, these results support provenance as a real downstream application, but not as a lineage-only or universal ancestry oracle: the downstream fingerprint is multi-factor, not purely genealogical.

Table 21: Supporting provenance audits.

Audit	Key numbers	Consequence
Official vs. unsloth audit	Pooled split 0.953629. HaluBench 0.959167. Align-Wrinkle 0.755000. MemProxy 0.691667.	Strong lineage-like packaging split; downstream use.
OOD provenance audit	In-domain 0.9516 / 0.9951. Matched-pair transfer stays centered; no saturation.	More specific than a broad OOD detector.
Laundering robustness	Utility 0.5839 vs. 0.5705. Fingerprint targets stay at 1.0; controls are much lower.	Utility can be preserved while the fingerprint survives.
Row-local trajectory	Full 0.954435 / 0.995542. Shape-only 0.954032 / 0.995592. Endpoint-only 0.953629 / 0.995532.	Robust to path compression; endpoints suffice.
Same-family scaling control	`model_tag` 1.000, even with `depth_only`. Controls: benchmark 0.829, prompt-group 0.558, label 0.573.	Scale is also a strong fingerprint, so provenance stays downstream.

C.11 Benchmark-Specific Factuality Anchors

The HaluBench source-split studies are useful calibration views, but they remain coarse and do not define a benchmark-global factuality axis. The TruthfulQA second-anchor experiment provides a second factuality check, but the signal remains benchmark-specific and can flip orientation under the default FAIL-oriented sign convention. The BoolQ third-anchor bundle sits between HaluBench and TruthfulQA: it preserves usable signal, but again fails to recover a geometry-only detector and therefore reinforces the bounded reading rather than restoring a shared truth axis.

C.12 Supplementary Control Audits

These supplementary audits test several tempting overreads without changing the paper's headline. First, supervised semantic-lexical partitions in the

tested residual runner do not create a distinct CRJ-fidelity regime over single-global, raw-k-means, or random-hash controls. Second, attention-state and attention–MLP-concat proxies are mixed rather than uniformly stronger: corridor rows remain inside the existing envelope, Qwen late changes little, and only late Llama shows a modest lift. Third, a bounded depth-component audit sharpens the fragility reading without producing a simple reversal mechanism: dominant-PC flattening does not restore the local law, PC sharpening hurts, and several matched gain/rebalance controls also degrade fidelity. A corrected three-seed repeat preserves the same interpretation on the principal rows, with stable corridor fragility under dominant-PC perturbation and no positive late-collapse rescue row for `rebalance_to_secondary`. Fourth, whole-block context-containing surrogate-to-differential calibrations are strong in Qwen and Llama, but attention-branch and cross-token extensions remain weak even after the low-rank bridge controls, whose best-of-family exact-JVP gains are too small to overturn the main boundary. Fifth, adjacent-target audits keep the factuality target split explicit: confidence is distinct from the tested truth/fail split, answer faithfulness remains bounded, and evaluator-disagreement targets stay mixed. Finally, a second-slice semantic stability check preserves only a narrow kinship-only stable subset and keeps continuation positives at zero. Taken together, these controls close remaining interpretive loopholes without changing the main claim set.

C.13 Branch-Selective Corridor Controls

Additional structural controls sharpen the depth account beyond the original whole-layer correlations. Under shared partitions, branch-local fits show that Qwen, Llama, and Pythia are MLP-led in the middle-layer corridor, while OPT is attention-led. Late layers then erase, flatten, or reverse that dominance. A light norm-preserving branch-scaling check moves block fidelity in the predicted direction for Llama and OPT, and a matched raw-scaling pass extends the same sign-flip pattern to Pythia, while Qwen remains comparatively stable under both perturbation styles. The structural summary is therefore branch-selective and family-dependent: the reusable regularity is a mid-depth corridor followed by a relatively stable late-collapse band, not a universal peak-layer law and not a gated-only theorem.

C.14 Depth Occupancy Controls

The most direct density-based attack on the depth result is that the main K=64 layer sweeps might simply be under-occupied local fits. The occupancy audit makes that narrow explanation difficult to sustain. In the three main depth sweeps, the late-collapse bands are not lower-occupancy bands: median train occupancy changes only from 14.0 $\rightarrow$ 14.5 in Qwen, 14.0 $\rightarrow$ 13.5 in Llama, and 8.0 $\rightarrow$ 8.5 in Phi-3 when moving from the middle corridor to the late-collapse band. Layer-wise train-count correlations are non-negative in all three sweeps, so later layers are not merely becoming emptier. Separately, an occupancy-controlled K-sweep on Qwen layer 12 keeps occupancy near 44–50 samples per cluster and shows held-out MSE improving rather than degrading as K increases. These controls do not prove occupancy irrelevance, but they make the U-shape much harder to dismiss as a simple low-occupancy local-fit artifact.

C.15 Split-Hygiene Generalization Audit

The 18-arm generalization audit directly tests whether the corridor ordering is an artifact of the official held-out split or of empty clusters. The answer is mostly negative. Official and within-train means are almost identical: overall cosine is 0.1876 versus 0.1891, and corridor cosine is 0.3330 versus 0.3339. Mean train–test gaps stay near zero, and cluster-level gaps do not show train inflation. Occupancy thresholding also leaves the regime means essentially unchanged from threshold 1 to threshold 30. The boundedness clause is equally important: Qwen and Llama carry the clear corridor signal in this audit, while the original Phi-3 anchors are near floor under this particular protocol. A dedicated Phi anchor sweep then recovers layers 10/11/16/17 and keeps 27/31 low, so the stricter Phi result is anchor/protocol-sensitive rather than uniformly negative. This audit therefore supports the split/occupancy robustness of the regime map but does not promote the result to a universal all-family peak-layer law.

C.16 Partition-Family Audit

A 120-arm partition-family audit asks whether the same regime readout depends on raw Euclidean k-means. For Qwen-3.5-4B, it does not appear to. Across raw, cosine, diagonal-whitened, PCA-whitened/Mahalanobis-proxy, gate-space, gate-cosine, MLP-boundary-proxy, and

random-hash partitions, the corridor anchor remains tightly grouped: structured-partition mean cosine is 0.4805 under the official protocol and 0.4817 under the within-train protocol, with ranges 0.0023 and 0.0019. The late-collapse anchor remains low under the same sweep (0.0362 and 0.0348), again with small structured-partition ranges. This rules out the narrow objection that the Qwen depth profile is created only by raw Euclidean k-means.

The same audit also narrows the interpretation. The unpartitioned single-global control is close to raw k-means at the Qwen corridor anchor, and the random-hash control is also close. A dedicated Phi replication shows that layers 10/11/16/17 likewise remain non-floor across metric, gate-space, and boundary-proxy partitions, while layers 27/31 remain low and layer 12 remains split-sensitive. The resulting reading is that, under this operator-surrogate protocol, the corridor-versus-collapse profile is partition-stable for matched model/layer anchors, while Phi-3 exposes anchor/protocol sensitivity and the null controls block a semantic-cluster overread.

A learned-feature check extends this boundary without closing it. Using a public Qwen-3.5-4B layer-18 TopK SAE, sign-hash partitions over selected reasoning, GSM8K-contrastive, norm-top, and random SAE features were often severely imbalanced. An occupancy-balanced comparison therefore clustered the same SAE encoder preactivations directly with K=32. In that cleaner comparison, raw k-means reached test cosine 0.3989/0.3962 under official/within-train protocols, while SAE reasoning, SAE GSM8K, and SAE random-feature k-means stayed at 0.3967/0.3952, 0.3967/0.3960, and 0.3964/0.3954, respectively. The structured range was only 0.0024 official and 0.0011 within-train. A hook-alignment diagnostic explains why this remains a boundary check rather than a closure theorem: the SAE reconstructs its intended residual post-layer-18 hook much better than the MLP-input hook used by the CRJ target (variance explained 0.881 versus about 0.71). Thus learned-feature partitions do not recover semantic partition specificity in this setup. Exact-hook same-layer transcoder features, exact gate activation-pattern stable-region proxies, compact attention-state partitions, token-shape supervised partitions, and a first shifted-corpus check give the same bounded reading, while stronger task-label/semantic-label supervision, full attention/KV-state regions, larger corpus-shift suites, and residual/full-block SAE-aligned targets remain open.

C.17 Exhaustive Layer Tables for the Primary Depth Sweeps

Because some parts of the appendix were already functioning like a layer ledger, we make that ledger explicit here. Tables 23–25 report every analyzed layer for the full-layer primary depth sweeps, rather than only the paper-facing regime anchors. These rows are exhaustive for the main depth evidence used to support the corridor-versus-collapse reading: Qwen and Phi-3 were swept through layer 31, while Llama was swept through layer 27. The point of exposing the full tables is not to enlarge the claim. It is to make the regime shape auditable without forcing the reader to reconstruct it from a mix of anchor rows and auxiliary CSV tables.

C.18 Hidden-State Access Family Extension

An access-family audit sharpens the factuality boundary considerably. The original raw-state linear baselines remain weak, but selected-layer hidden-state probes recover substantial grouped signal on Qwen and Llama across HaluBench, TruthfulQA, and BoolQ, including harder source-, pair-, and length-style holdouts. The strongest grouped results reach 0.826 on Qwen HaluBench and 0.786 on Llama TruthfulQA, while the more conservative representative rows reported in the main text remain lower. At the same time, score-free geometry remains near chance even after adding richer geometry, trajectory, and norm summaries. A grouped detector-ladder audit further shows that the matched grouped geometry deltas are uniformly negative, not merely noisy. The resulting picture is therefore a three-way split: weak raw-state linear access, materially stronger hidden-state access, and still-weak pure score-free geometry, with CRJ geometry belonging primarily to the diagnostic boundary rather than to a strong additive detector claim.

C.19 Alignment-Sensitive Transfer

The raw-coordinate cross-model detector family does not transfer. However, prompt-matched Procrustes alignment partially restores probe-rich transfer, while geometry-only and trajectory-only families stay at chance. This makes the transfer result more specific than a coordinate-free private-geometry claim. Alignment matters, but it does not recover a strong score-free geometry account.

C.20 Exhaustive Layer Tables for the Primary HaluBench Sweep

The same clarification matters for the factuality appendix. Tables 36–38 report every analyzed layer for the full-layer HaluBench raw-state linear probe sweeps, so readers can inspect the actual layerwise trajectories rather than only the paper-facing anchor summaries. The same full-layer treatment extends to the secondary anchors as well. Those full-layer sweeps do not recover a shared truth axis: Qwen TruthfulQA stays above chance across layers (best AUC ≈ 0.587 at layer 0), Llama TruthfulQA stays below chance throughout (best AUC ≈ 0.408 at layer 27), Llama BoolQ remains above chance (best AUC ≈ 0.610 at layer 0), and Qwen BoolQ remains below chance (best AUC ≈ 0.424 at layer 0). The exhaustive tables therefore strengthen the existing bounded reading rather than change it: the raw-state signal is benchmark- and model-specific even before any transfer question is asked. The corresponding exhaustive tables are collected together at the end of the appendix so that they can be consulted as a ledger without interrupting the prose flow.

C.21 Broader Concept-Family Controls

The broader concept-family extension broadens the family coverage beyond the initial family/relational cluster, but it still does not collapse into a universal dial. It is best read as a coverage extension that confirms the intervention result generalizes to additional clusters while remaining concept-specific.

C.22 Local Parameter-Edit Linearization Audit

The local-edit audit separates a hidden-response question from a token-facing question. Across 288 rows, hidden-response local linearization is strong: mean cosine is 0.9892 and median cosine is 0.9962. Tracked log-probability response is much weaker: mean cosine is 0.5586, median cosine is 0.6303, and mean relative L2 is 1.0796. The gap is visible in every model family, with tracked-response cosine highest in Qwen (0.7432), lower in Llama (0.5687), and weakest in Phi-3 (0.3640), while hidden-response cosine stays near 0.99 throughout. The accompanying token-family audit gives only modest semantic separation: mean positive/boundary absolute ratio is 1.3151, and sign-level fractions remain near balanced. This audit supports the central semantic boundary: local hidden dynamics are often linear enough to predict, but token-facing causal specificity is a harder and more heterogeneous target.

C.23 Local Conditional Intervention Audit

A constructive intervention audit asks whether the failure mode belongs to intervention in general or specifically to global unconditional directions. On Qwen layer 12 cluster 27, local conditional policies preserve the target boundary effect while materially reducing control spillover: the magnitude of the target-versus-control boundary gap increases from roughly 0.00078 under `global_target` to 0.00849 under `conditional_target_only`, and routed-cluster policies also improve contrast-level specificity. Two additional family-oriented Qwen target clusters at the same layer extend that constructive read: for clusters 1814 and 1592, `routed_cluster_v1` yields the best specificity-positive gaps (0.001058 and 0.000710) and specificity-boundary gaps (0.000300 and 0.000381), both above the corresponding global controls. A matched Llama family arm remains weak, so the constructive conclusion is replicated but family-sensitive rather than universal.

C.24 Sequence-Level Semantic Audit

A frozen-casebook rollout tests whether this bounded constructive signal survives beyond the next token. It does not, at least not in a strong continuation-level sense. Across Qwen clusters 27/1814/1592 and the matched Llama arm, continuation-level target-family and control-family hard-hit rates both remain at zero, so the realized sequence-level specificity gap does not improve. The responsive Qwen clusters still retain weak soft target-mass shifts under rollout—for example, cluster 1592 shows a best positive-mass delta of about $+2.1 \times 10^{-5}$ and cluster 1814 reaches about $+3.1 \times 10^{-5}$—but those shifts are too small to overcome the model's default continuation mode. The constructive semantic reading is therefore trust-region local: local conditional policies can improve next-token specificity in responsive families, but the present evidence does not yet show robust continuation-level semantic control.

The multiseed continuation audit probes whether this negative rollout result was too dependent on the original target clusters or too weakly seeded. Using five seeds, longer rollouts, and a broader policy menu, it still finds only a nar-

row Llama-3.2-3B continuation pocket: `kinship` and `temporal` under `mean_difference` / `global_mean_difference` reach grouped continuation pass rate 1/3, but the mean target rollout-positive rate is only 0.0278 and the maximum remains 0.0833. The matched Qwen-3.5-4B corridor arm yields no surviving continuation-positive row. This keeps the main ambiguity bounded in the conservative direction: continuation-level positives are not literally absent, but they remain too sparse and weak to overturn the next-token-local reading.

C.25 Fixed-Protocol Family Sweep

A fixed-selector benchmark asks whether the local-policy reading survives once both family selection and policy choice are frozen in advance. On Qwen layer 12, we evaluate five lexical families (`kinship`, `occupation`, `place_identity`, `sentiment`, and `negation`) using one fixed selector, one fixed control rule, and one fixed policy menu. The resulting picture is modest but informative. Local policies retain positive best next-token specificity in four of the five families and outperform the best global policy in three of the five, with the clearest bounded signals appearing in `kinship` and `negation`. However, the margins are small, `sentiment` does not replicate prospectively, and all five families remain `soft_only` under continuation evaluation. The selector also still prefers extremely high-lift lexical clusters against near-zero-positive controls, so this benchmark is prospective but not yet a broad concept-general steering evaluation. The conclusion is therefore narrow: the Qwen cluster examples above are not a one-cluster artifact, but the constructive gain remains weak, lexical, and next-token local rather than a benchmarked semantic steering law.

A broader eight-family fixed-protocol benchmark pushes the same setup to `kinship`, `occupation`, `place_identity`, `sentiment`, `negation`, `temporal`, `modality`, and `quantity`. It preserves the same ceiling rather than upgrading the claim: local policies remain weakly positive at next-token specificity in some families, beat the best global policy in a majority but not all of them, and still never produce phrase-level continuation hits. All eight families remain `soft_only`, so broadening the concept bank increases coverage but still does not turn the local-policy result into a sequence-level steering law.

C.26 Prospective Layer-Selector Benchmark

A cross-layer benchmark asks a different operational question: if family selection, policy menu, and continuation evaluation are all frozen, does the middle-layer corridor itself behave like a prospective best layer for semantic intervention? On Qwen we compare one pre-corridor layer (L2), one in-corridor layer (L12), and one late-collapse layer (L27) on the same five lexical families. The result is mixed rather than decisive. The corridor yields the largest mean absolute local next-token specificity gap (3.22×10^{-4}, versus 6.99×10^{-5} for the pre-corridor arm and 2.88×10^{-5} for late collapse), but it is not the clean winner by matched-global lift: pre-corridor layers beat their matched global baselines in all five families, corridor layers in three, and late-collapse layers in two. Per-family winners are similarly split: `kinship` is strongest in the corridor, `occupation` and `sentiment` favor the pre-corridor arm, and `place_identity` is split between late-collapse absolute gap and corridor matched-global lift. All continuation reads remain `soft_only`. The operational lesson is therefore not that the corridor is a universal best layer, but that late-collapse is an unattractive default and that the corridor marks a promising semantic regime whose usefulness still depends on family and metric.

C.27 Corridor Law and Trust-Region Use

Retrospective synthesis across these supplementary audits, now tempered by the prospective layer-selector benchmark, suggests two narrow operational takeaways. First, relative peak depth is too loose to support a canonical best-layer law, whereas late-collapse onset is substantially tighter across the tested families. Second, the corridor gives a better regime prior for semantic intervention than for downstream factuality readout selection, but only in a narrow sense: all reported semantic checks lie inside it, while five of six benchmark-specific hidden-state factuality optima lie outside it and two sit at or beyond collapse onset; meanwhile the prospective benchmark shows stronger absolute lexical effects in the corridor, mixed matched-global lift, and no continuation-level wins. This is why we treat Flash-Jacobian as a regime and trust-region diagnostic for linear semantic intervention rather than as a universal best-layer oracle.

The broader eight-family fixed-protocol benchmark reinforces the same bound from a different angle. Even after widening the family bank, the

fixed selector still only preserves weak next-token-local gains in some families, the best global policy remains competitive in others, and phrase-level continuation hits remain zero across all eight families. The broader benchmark therefore strengthens the claim that FLASH-JACOBIAN is a narrow-regime diagnostic, not a continuation-level steering oracle.

C.28 Routing Stability Audits

The routing audits identify a few narrow slices where geometry-aware perturbations beat same-norm random baselines by substantial margins, so the intervention line is not dead. However, the corresponding seed-stable repeat evaluations also show that the candidate centroid directions remain nearly perfectly collapsed. A retrospective direction-equivalence audit gives the same warning from another angle: low-alignment non-V_1 directions often reproduce a large fraction of the V_1 token-facing magnitude, and in 42.9% of audited target/alpha groups they exceed the V_1 absolute NLL effect. The correct interpretation is therefore functional and reproducible routing sensitivity, not a uniquely isolated causal mechanism.

C.29 Experimental Details

C.30 Hardware Configuration

Primary operator-fitting measurements and many mainline LLM runs were conducted on Google Colab using NVIDIA L4 GPUs with the following representative specifications: GPU: NVIDIA L4 (23.7 GB VRAM, Compute Capability 8.9), system RAM: 53.0 GB, storage: 112.6 GB SSD, software: PyTorch + CUDA (standard transformer stack) with bundled Flash-KMeans components. Some later closure and transfer follow-up packages cited in this paper were executed on other GPU hosts, including mixed RunPod V100-, RTX-4090-, RTX-5080-, RTX-PRO-4500-, A100-, and L40-class environments; the relevant package-local analysis notes record those execution environments explicitly.

Synthetic scalability experiments were conducted on CPU (Apple MacBook Air (M1, 2020; 16 GB RAM)).

C.31 Hyperparameters

C.32 Data Collection

Hidden states were extracted from WikiText-2 (Merity et al., 2017) in three steps. We first sampled 1,000 passages with an average length of roughly 512 tokens. We then ran the model up to the target layer ℓ and recorded the paired hidden states (X_ℓ, Y_ℓ). Finally, we uniformly subsampled the pooled states down to N=100,000 pairs to keep fitting tractable.

C.33 Memory and Complexity

C.34 Synthetic Scalability Validation

Before deploying FLASH-JACOBIAN on large-scale language models, we first validated its fundamental scalability limits using synthetic datasets on a standard consumer CPU. This allowed us to cleanly isolate the $O(\text{tile}\times K)$ memory behavior of FlashAssign from GPU-specific overheads. The synthetic data contained random normally-distributed clusters in d=512.

As shown in Table 30, scaling both the number of clusters K and the dataset size N by 10× (from Small to Large) would make a standard $N\times K$ distance matrix impractical. FLASH-JACOBIAN still scales close to linearly in wall-clock time (61s to 591s), which is consistent with the blocked assignment design.

C.35 Resolution Sweep: The K–N Trade-off

A core hyperparameter governing local linear approximations is the resolution K (number of clusters). While higher K implies smaller, more localized regions where the linear assumption holds better, it also fragments the dataset, leaving fewer samples n_k per cluster.

Table 31 illustrates this trade-off. Under a fixed sample budget (N=30k), increasing K eventually hurts test accuracy because the average occupancy $n_k = N/K$ becomes too small. When N is scaled proportionally with K to maintain constant average occupancy (right column: N=80K), test MSE improves steadily as K increases. High-resolution partitions therefore require correspondingly large datasets.

C.36 LLM Baselines and Ablations

We validated the real-world operational effects of K and adapter rank r by ablating them directly on Qwen-3.5-4B Layer 12, shown in Table 32.

Table 32: Qwen-3.5-4B Layer 12 baselines and ablations (N_{train}=80k, FP16). Top: resolution sweep. Bottom: LoRA rank ablation at K=2000.

K	r	MSE(y) ↓	MSE($\hat{y} = x$) ↓	Improv.(%) ↑	$R^2(y)$ ↑	Time (s)
1	16	0.00272	0.00429	36.5	0.934	11.1
250	16	0.00229	0.00429	46.5	0.944	36.1
500	16	0.00217	0.00429	49.4	0.947	49.6
1000	16	0.00215	0.00429	49.9	0.948	84.0
2000	16	0.00213	0.00429	50.4	0.948	154.4
2000	1	0.00236	0.00429	45.0	0.943	153.0
2000	2	0.00231	0.00429	46.2	0.944	155.2
2000	4	0.00225	0.00429	47.6	0.945	156.0
2000	8	0.00218	0.00429	49.1	0.947	154.5
2000	16	0.00213	0.00429	50.4	0.948	155.0

The top half shows the effect of partitioning the manifold. A single global operator (K=1) improves MSE by 36.5% over the identity mapping, while splitting the space into K=2000 local regions raises that improvement to 50.4%, indicating that local partitions capture variation that a single global frame misses. The bottom half shows the sensitivity to the LoRA bottleneck rank r. As r increases from 1 to 16, predictive fidelity improves modestly, which indicates that the residual dynamics can use additional rank when the data support it. Time is nearly unchanged across r because the Woodbury update removes the dominant $d \times d$ inverse from the inner loop.

The convergence properties of the Alternating Least Squares (ALS) engine within Flash-Jacobian are well-behaved on this setup. Figure 3 shows the fit-only loss trajectory for Qwen-3.5-4B Layer 12 with K=2000 clusters over 20 iterations. The system exhibits rapid initial convergence, with the dominant reduction occurring in the first five alternating iterations and the loss flattening thereafter. We terminate early after 20 iterations, requiring only 156.5 seconds overall, indicating that standard local optimization methods are practical over the extracted LLM representations.

C.37 Robustness to Initialization Seed

To verify that the fitted partitions are not overly sensitive to centroid initialization, we repeated the full pipeline (Qwen-3.5-4B Layer 12, K=2000, r=16, N=80k) three times with independent random seeds. The resulting test MSEs were 0.002131, 0.002198, and 0.002167, for a mean of 0.002165 and a standard deviation of 3.36×10^{-5}. Mean test cosine fidelity similarly stayed tight at 0.501, 0.498, and 0.500 ($\sigma = 1.5 \times 10^{-3}$). These aggregate metrics indicate that the fitted piecewise operators are stable under initialization noise.

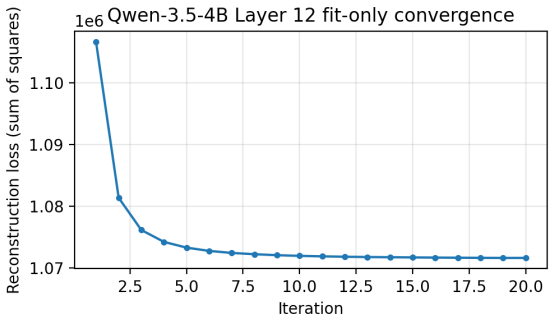

Figure 3: Fit-only loss convergence for Qwen-3.5-4B Layer 12 (K=2000, N=98,412). The loss drops quickly in the first few iterations and then flattens. Total fit time: 156.5s on NVIDIA L4 (Google Colab).

C.38 Rank-1 Dominance and Sample Occupancy

To characterize the internal shape of local operators, we inspect the right-singular vectors of each residual operator matrix B_k. Many clusters exhibit a moderately dominant leading component, but the singular value ratios also reveal a sample-size artifact. Across the full population of K=2000 clusters, 648 clusters (32.4%) satisfy $\sigma_1/\sigma_2 > 2$, and 226 clusters (11.3%) satisfy $\sigma_1/\sigma_2 > 5$.

However, the clusters with the most extreme rank-1 dominance ($\sigma_1/\sigma_2 > 10$) are disproportionately low-occupancy clusters with fewer than 10 training samples. This suggests that the strongest apparent one-dimensional structure is often a low-sample artifact rather than a stable local concept. Clusters with healthier occupancy (> 20 samples) more often retain a dominant first singular vector together with a non-trivial secondary subspace.

C.39 Differential Jacobian Fidelity on the MLP Submodule

This appendix clarifies the relationship between fitted empirical input–output surrogates and exact pointwise differential Jacobians. We focus on settings where position-wise JVPs are well defined: a toy ReLU MLP and the isolated MLP/FFN submodule inside a transformer block. We compare predicted vectors against exact autograd JVPs on held-out test points using cosine similarity, relative ℓ_2 error, and norm ratio.

We consider region-constant operators of the form $\widehat{J}_k = \mathrm{Diag}(d_k) + A_k B_k$, including diagonal-only baselines. Pure low-rank families tend to collapse toward near-zero predictions whenever the true Jacobian retains substantial full-rank residual structure.

The toy ReLU MLP results expose the basic partitioning limit. For a 2-layer ReLU MLP with d=64, an oracle regime-aligned partition recovers the exact Jacobians by construction (cosine = 1). But the number of distinct activation patterns grows rapidly toward $K \approx N$, so exact structural partitioning is not practical at realistic scale (Table 33).

On Qwen-3.5-4B Layer 12 MLP (d=2560), simple diagonal operators still reach cosine similarity ≈ 0.48 with relative $\ell_2 \approx 0.88$ (Table 34). Adding LoRA does not improve on the diagonal-only baseline, which suggests that the diagonal component captures much of the recoverable pointwise MLP-Jacobian structure in this setup.

A surrogate-to-differential calibration extends this check beyond a single Qwen corridor layer. For Qwen-3.5-4B, Llama-3.2-3B, and Phi-3-mini, we fit four diagonal operators on matched isolated-MLP anchors: an exact-JVP fit, a centered input–output fit, an in-cluster secant fit, and a shuffled-JVP null. Each fitted operator is then evaluated on exact held-out JVPs, centered held-out IO residuals, and held-out in-cluster secants. A full-depth calibration pass then extends the same calibration across all layers for all three models. The goal is not to replace CRJs with pointwise Jacobians, but to validate the naming boundary by measuring when the population surrogate and exact differential estimands coincide and when they separate.

The full-depth sweep keeps the same qualitative shape while removing the old anchor-only caveat. Under the multi-seed exact-JVP anchor calibration, Qwen peaks at layer 7 (cosine 0.564 ± 0.004, shuffled-null margin 0.526), Llama at layer 14 (cosine 0.471 ± 0.004, margin 0.459), and Phi-3 at layer 11 (cosine 0.317 ± 0.003, margin 0.316). Early and late layers remain materially weaker, and finite-IO-fitted operators can still retain some nonzero exact-JVP overlap even after the main corridor has decayed. This supports the terminology used throughout the paper: CRJs are conditional/operator surrogates whose relation to exact pointwise differentials must be calibrated, not assumed, and the calibration rows are a validation target rather than the main CRJ estimand. The corresponding exhaustive layer tables are collected together at the end of the appendix so that the calibration ledger stays contiguous without interrupting the argument.

To isolate density effects, we also ran an occupancy-controlled sweep. Scaling n_{train} with K to maintain an occupancy ratio of roughly 32 keeps JVP fidelity fairly stable (cosine $\approx$ 0.47–0.48) across $K \in \{32, 64, 128, 256\}$. This suggests that the degradation under fixed n_{train} is driven largely by insufficient cluster occupancy.

Finally, we tested activation-boundary-aware partitions to see whether more structure-aware clustering could break this plateau. Importance-weighted sign hashes, boundary ternary bins, and focused MiniBatchKMeans in the gated feature space all yield similar global JVP fidelity (cosine $\approx$ 0.48) to the standard x-KMeans baseline. Moderate partition refinements alone therefore do not push fidelity beyond the observed plateau.

C.40 Exhaustive Layer Ledgers

This closing section collects the full-layer appendix tables that support the bounded factuality and isolated-MLP calibration readings. They are grouped here deliberately so that the appendix prose can run continuously while the exhaustive ledgers remain easy to inspect in one place.

Table 22: Supporting regime-map and supplementary control audits.

Audit	Key numbers	Consequence
Boundary tightening checks	Supervised semantic partitions do not beat single-global / raw-k-means / random-hash. Attention-state proxies are mixed; late Llama only shows a modest lift. Whole-block calibrations are strong, while attention-branch and cross-token arms stay weak.	Tightens multiple overreads without expanding the headline claim.
Branch corridor	Qwen/Llama/Pythia are MLP-led in corridor layers; OPT is attention-led. Late layers erase, flatten, or reverse branch dominance.	Supports corridor-plus-late-collapse, but as branch-selective and family-dependent rather than universal.
Occupancy control	Median train occupancy shifts only: 14.0 → 14.5 (Qwen), 14.0 → 13.5 (Llama), 8.0 → 8.5 (Phi-3). Qwen layer-12 occupancy-controlled K-sweep keeps ≈ 44–50 samples per cluster and improves held-out MSE as K rises.	Makes the depth U-shape hard to dismiss as a simple under-occupied local-fit artifact.
Split-hygiene audit	Overall cosine: 0.1876 official vs. 0.1891 within-train. Corridor cosine: 0.3330 vs. 0.3339. Thresholding from 1 to 30 leaves regime means nearly unchanged.	Supports split/occupancy robustness while keeping Phi-3 anchor sensitivity explicit.
Partition audit	Qwen structured-partition corridor mean: 0.4805 official, 0.4817 within-train; ranges 0.0023/0.0019. Late-collapse mean: 0.0362/0.0348.	Rules out a raw-Euclidean-only explanation, but null controls still block a semantic-atlas overclaim.

Table 23: All analyzed Qwen layers for the primary depth sweep (`kmeans_x` / `jvp_diag_fit`, K=64). Cells report cosine / relative ℓ_2.

L	Cos	Rel ℓ_2	L	Cos	Rel ℓ_2
0	0.171	0.986	16	0.408	0.912
1	0.168	0.988	17	0.410	0.911
2	0.240	0.971	18	0.394	0.918
3	0.242	0.970	19	0.424	0.905
4	0.280	0.960	20	0.321	0.946
5	0.433	0.900	21	0.223	0.976
6	0.500	0.865	22	0.260	0.966
7	**0.557**	0.829	23	0.205	0.980
8	0.514	0.856	24	0.140	0.993
9	0.461	0.886	25	0.121	0.996
10	0.448	0.893	26	0.104	0.999
11	0.490	0.871	27	0.023	1.008
12	0.473	0.881	28	0.011	1.008
13	0.450	0.892	29	0.040	1.006
14	0.382	0.924	30	0.047	1.005
15	0.399	0.916	31	0.079	1.002

Table 24: All analyzed Llama layers for the primary depth sweep (`kmeans_x` / `jvp_diag_fit`, K=64). Cells report cosine / relative ℓ_2.

L	Cos	Rel ℓ_2	L	Cos	Rel ℓ_2
0	0.069	1.004	14	**0.499**	0.864
1	0.149	0.990	15	0.455	0.889
2	0.191	0.982	16	0.366	0.931
3	0.278	0.960	17	0.314	0.950
4	0.440	0.895	18	0.269	0.964
5	0.452	0.889	19	0.247	0.970
6	0.476	0.876	20	0.143	0.992
7	0.457	0.887	21	0.068	1.003
8	0.465	0.883	22	0.026	1.007
9	0.463	0.884	23	0.005	1.008
10	0.455	0.888	24	0.009	1.008
11	0.464	0.884	25	0.032	1.006
12	0.477	0.876	26	0.075	1.001
13	0.495	0.867	27	0.123	0.994

Table 25: All analyzed Phi-3 layers for the primary depth sweep (`kmeans_x` / `jvp_diag_fit`, K=64). Cells report cosine / relative ℓ_2.

L	Cos	Rel ℓ_2	L	Cos	Rel ℓ_2
0	0.068	1.003	16	0.241	0.971
1	0.027	1.008	17	0.230	0.974
2	0.051	1.006	18	0.213	0.978
3	0.063	1.003	19	0.191	0.983
4	0.083	1.000	20	0.142	0.992
5	0.098	0.999	21	0.100	0.999
6	0.159	0.988	22	0.064	1.003
7	0.188	0.982	23	0.033	1.007
8	0.230	0.972	24	0.020	1.008
9	0.241	0.969	25	0.011	1.009
10	0.259	0.965	26	0.011	1.008
11	**0.272**	0.961	27	0.010	1.009
12	–	–	28	0.011	1.009
13	0.243	0.970	29	0.008	1.012
14	0.261	0.965	30	0.005	1.009
15	0.262	0.965	31	0.013	1.009

Table 26: Supporting factuality audits.

Audit	Key numbers	Consequence
Leakage ablation	`core_minus_score`: RF 0.9516 / 0.9999. `core_minus_direct`: RF 0.5029, HGB 0.5069. Geometry-only / trajectory-only ≈ chance.	Composite detector; probe-derived channels matter.
Hidden-state access	Grouped best rows: Qwen HaluBench 0.775 / 0.712; Llama TruthfulQA 0.772 / 0.687; BoolQ both ≈ 0.63.	Raw-state linear is too weak; grouped hidden-state access is not uniformly weak.
Score-free geometry	Pooled grouped best ≈ 0.505. Best within-model grouped 0.5304 / 0.5444. All `geom_delta_auc` values are negative (-0.2845 to -0.4499).	Pure score-free geometry still does not recover factuality.
Geometry delta	Best `core_geom`: 0.5827.	CRJ geometry does not earn a strong additive claim on top of the core.
Perm. sanity	Permuted AUC: hidden ≈ 0.51; core ≈ 0.53.	Rules out simple grouped-split leakage.
TruthfulQA anchor	Qwen full 0.576 / 0.570. Llama full 0.368 / 0.389; sign-flipped ≈ 0.632.	Second anchor; benchmark-specific and orientation-sensitive.
BoolQ third anchor	Qwen oriented 0.560 / 0.531. Llama 0.619 / 0.617. `core_minus_direct` remains near chance.	Third anchor; bounded factuality reading still holds.
Adjacent target	High-confidence target (q75) reaches ≈ 0.90 and up to 0.98 on HaluEvalQA. Best overconfident-error target (q75) AUCs: 0.919 (Llama TruthfulQA), 0.855 (Llama HaluEvalQA), 0.830 (Llama HaluBench), 0.743 (Qwen TruthfulQA).	Confidence is a distinct adjacent target; overconfident-error access is real but bounded and benchmark-sensitive.
Aligned transfer	Prompt-matched Procrustes mean oriented AUC ≈ 0.866 for probe-rich families. Geometry-only / trajectory-only stay at chance. Best in-template AUC 0.9992. Mean cross-template AUC drops range from about -0.02 to -0.23. Worst single drop is about -0.42.	Transfer failure is raw-coordinate- and detector-family-specific.
Template sensitivity	Worst `model_incorrect` agreement: 0.110 (Llama TruthfulQA).	Within-template hidden-state access remains real, but transfer degrades enough to block a template-global truth-geometry reading.

Table 27: Supporting intervention audits.

Audit	Key numbers	Consequence
Descriptive vs. causal audit	Magnitude correlation $\rho \approx 0.62$. Specificity correlations $\rho \approx 0.15$–0.26.	Descriptive quality tracks size better than specificity (132 non-random directions: 63 target-principal, 69 target-like).
Broader concept control	cluster39: `NLL` $\Delta = +0.040$, contrast $\Delta = -0.0006$, top-1 0.0. Family/digit/case controls stay concept-specific.	Coverage extension, not a universal dial.
Local conditional audit	Qwen cluster 27: boundary-gap magnitude 0.00849 (`conditional`) vs. 0.00078 (`global`). Qwen clusters 1814/1592: best routed positive gaps 0.001058 / 0.000710 and boundary gaps 0.000300 / 0.000381. Llama stays at 10^{-5}–10^{-4}.	Constructive and replicated within a responsive Qwen family, but still family-sensitive and policy-dependent.
Sequence-level audit	Across Qwen clusters 27/1814/1592 and the matched Llama comparison, target/control hard-hit rates stay at 0.0. Qwen soft positive-mass deltas persist only weakly (best $\approx 2.1\times10^{-5}$ and 3.1×10^{-5} on clusters 1592/1814).	Bounds the constructive reading to a trust-region / next-token-local result rather than a sequence-level steering win.
Fixed-protocol family sweep	Five lexical families on Qwen layer 12. Positive next-token local specificity in 4/5 families; local policy beats the best global policy in 3/5; all five remain `soft_only` under continuation.	Weak prospective replication of the local-policy effect under a fixed selector, but still lexical, selector-driven, and not a continuation-level steering benchmark.
Prospective layer selector	Qwen layers L2/L12/L27. Mean local positive gap is strongest in corridor (3.22×10^{-4}), but matched-global lift is strongest in pre-corridor (1.43×10^{-5} versus 5.39×10^{-6} in corridor); late-collapse is weakest and all arms remain `soft_only`.	Mixed prospective support: useful for regime narrowing and avoiding late-collapse defaults, not for a clean best-layer law.
Narrow routing stability check	`changed_adv` 0.767–0.933. `abs_margin_adv` 2.436–4.212. Centroid collapse span 0.0.	Functional and reproducible, but not uniquely causal.

Table 28: Hyperparameters for all experiments.

Parameter	Value
LoRA rank (r)	16
Regularization (λ)	10^{-2}
Max iterations ($T_{\max}$)	20
Convergence tolerance (τ)	10^{-4}
Precision	FP16 (LLM), FP32 (synthetic)
Initialization	Xavier uniform (LoRA) Random sample (centroids)

Table 29: Memory comparison (FP16, K=40k, N=500k, d=3072, r=16). ''Naive'' stores full Jacobians and the full distance matrix; FLASH-JACOBIAN stores low-rank factors and uses tiled assignment.

Component	Naive	FLASH-JACOBIAN
Distance matrix ($N\times K$)	40.0 GB	(*not materialized*)
Jacobian storage (per-cluster)	755.0 GB†	7.9 GB‡
Data (X, Y)	6.1 GB	6.1 GB
Centroids ($\bar{X}, \bar{Y}$)	0.5 GB	0.5 GB
Total (lower bound)	**≈801 GB**	**≈14.5 GB**

†Full Jacobians: $K\times d^2\times 2$ bytes (FP16).

‡LoRA factors A, B: $2\times K\times d\times r\times 2$ bytes (FP16).

Table 30: Scalability validation on synthetic data (CPU). Time scaling is near-linear.

Scale	N	K_{fit}	Final Loss	Time (s)
Small	50k	1,000	1.80×10^8	61.3
Medium	200k	4,000	1.26×10^9	234.7
Large	500k	10,000	5.15×10^9	591.2

Table 31: Resolution sweeps on synthetic data. Fixed N degrades with increasing K; occupancy-controlled scaling (N=80K) improves.

K	Fixed N=30k (test MSE)	Occupancy-controlled N=80K (test MSE)
25	49.20	189.03
50	54.69	212.36
100	68.29	142.99
200	82.33	119.69
400	164.81	112.56
800	132.73	111.65
1600	138.13	103.77

Table 33: Toy ReLU MLP differential-Jacobian sanity checks (d=64).

Setting	K	Cosine↑	Rel. ℓ_2 ↓	Norm ratio↑	Rel. Fro↓
Regime-exact oracle (no overflow)	2001	1.000	0.000	1.000	0.000
Regime-exact oracle (cap=200)	201	0.513	1.010	1.033	1.003
x-KMeans + JVP-ridge	50	0.656	0.777	0.809	0.770

Table 34: MLP differential-JVP fidelity on Qwen-3.5-4B Layer 12 (n_{train}=1024, n_{test}=256).

Partition	Operator	Cosine↑	Rel. ℓ_2 ↓	Norm ratio↑
x-KMeans (K=64)	Low-rank ridge	0.096	0.997	0.150
x-KMeans (K=64)	Diagonal	0.475	0.880	0.479
x-KMeans (K=64)	Diagonal + LoRA	0.465	0.884	0.466

Table 35: Surrogate-to-differential calibration on isolated MLP anchors. Values are mean cosine on exact JVP, centered IO, and pair-secant evaluations.

Model	Layer	Regime	Fit	Exact JVP	Centered IO	Secant
Qwen	12	corridor	JVP diag	0.476	0.358	0.357
Qwen	12	corridor	centered-IO diag	0.467	0.372	0.365
Qwen	27	late	JVP diag	0.028	0.129	0.046
Qwen	27	late	centered-IO diag	0.012	0.225	0.118
Llama	14	corridor	JVP diag	0.513	0.278	0.251
Llama	14	corridor	centered-IO diag	0.484	0.314	0.285
Llama	27	late	JVP diag	0.106	0.136	0.117
Llama	27	late	centered-IO diag	0.046	0.224	0.142
Phi-3	11	corridor	JVP diag	0.292	0.126	0.126
Phi-3	11	corridor	centered-IO diag	0.236	0.139	0.131
Phi-3	31	late	JVP diag	0.005	0.010	0.002
Phi-3	31	late	centered-IO diag	0.014	0.153	0.173

Table 36: All analyzed Qwen layers for the full-layer HaluBench raw-state linear probe sweep. Cells report test AUC / test balanced accuracy.

L	AUC	Bal. acc.	L	AUC	Bal. acc.
0	0.465	0.489	16	0.502	0.492
1	0.470	0.486	17	0.494	0.497
2	0.447	0.486	18	0.504	0.494
3	0.452	0.481	19	0.478	0.489
4	0.458	0.481	20	0.481	0.483
5	0.467	0.478	21	0.468	0.481
6	0.485	0.483	22	0.482	0.481
7	0.473	0.483	23	**0.522**	0.486
8	0.473	0.483	24	0.504	0.486
9	0.471	0.486	25	0.483	0.483
10	0.481	0.486	26	0.502	0.492
11	0.471	0.486	27	0.471	0.494
12	0.475	0.489	28	0.471	0.478
13	0.475	0.489	29	0.480	0.503
14	0.477	0.489	30	0.483	0.497
15	0.483	0.489	31	0.486	0.506

Table 37: All analyzed Llama layers for the full-layer HaluBench raw-state linear probe sweep. Cells report test AUC / test balanced accuracy.

L	AUC	Bal. acc.	L	AUC	Bal. acc.
0	0.508	0.519	14	0.508	0.503
1	0.533	0.503	15	0.509	0.503
2	0.528	0.503	16	0.509	0.503
3	0.532	0.503	17	0.507	0.503
4	0.532	0.503	18	0.507	0.503
5	**0.534**	0.503	19	0.508	0.503
6	0.512	0.503	20	0.509	0.503
7	0.508	0.503	21	0.508	0.503
8	0.505	0.503	22	0.507	0.503
9	0.507	0.503	23	0.508	0.503
10	0.508	0.503	24	0.507	0.503
11	0.507	0.503	25	0.508	0.503
12	0.507	0.503	26	0.502	0.503
13	0.507	0.503	27	0.460	0.481

Table 38: All analyzed Phi-3 layers for the full-layer HaluBench raw-state linear probe sweep. Cells report test AUC / test balanced accuracy.

L	AUC	Bal. acc.	L	AUC	Bal. acc.
0	0.490	0.486	16	0.502	0.492
1	0.475	0.492	17	0.498	0.500
2	0.465	0.489	18	0.495	0.500
3	0.463	0.492	19	0.501	0.497
4	0.457	0.486	20	0.499	0.497
5	0.460	0.483	21	0.502	0.497
6	0.478	0.489	22	0.494	0.497
7	0.480	0.483	23	0.489	0.494
8	0.482	0.483	24	0.508	0.506
9	0.487	0.483	25	0.503	0.503
10	0.485	0.486	26	0.508	0.503
11	0.492	0.489	27	0.508	0.503
12	0.494	0.500	28	0.502	0.500
13	0.495	0.492	29	**0.509**	0.508
14	0.501	0.497	30	0.506	0.506
15	0.501	0.492	31	0.497	0.503

Table 39: All analyzed Qwen layers for the full-layer TruthfulQA raw-state linear probe sweep. Cells report test AUC / test balanced accuracy.

L	AUC	Bal. acc.	L	AUC	Bal. acc.
0	**0.587**	0.584	16	0.549	0.607
1	0.580	0.582	17	0.552	0.611
2	0.576	0.599	18	0.554	0.613
3	0.575	0.605	19	0.553	0.605
4	0.572	0.627	20	0.557	0.607
5	0.570	0.623	21	0.562	0.611
6	0.578	0.640	22	0.564	0.615
7	0.574	0.631	23	0.556	0.613
8	0.579	0.631	24	0.562	0.613
9	0.575	0.625	25	0.563	0.617
10	0.576	0.627	26	0.560	0.611
11	0.562	0.613	27	0.565	0.615
12	0.555	0.607	28	0.564	0.611
13	0.556	0.607	29	0.571	0.615
14	0.558	0.611	30	0.581	0.625
15	0.554	0.609	31	0.573	0.615

Table 40: All analyzed Llama layers for the full-layer TruthfulQA raw-state linear probe sweep. Cells report test AUC / test balanced accuracy.

L	AUC	Bal. acc.	L	AUC	Bal. acc.
0	0.407	0.413	14	0.366	0.385
1	0.392	0.405	15	0.351	0.373
2	0.384	0.393	16	0.354	0.371
3	0.365	0.373	17	0.362	0.385
4	0.374	0.387	18	0.370	0.393
5	0.375	0.389	19	0.365	0.387
6	0.378	0.389	20	0.384	0.401
7	0.369	0.385	21	0.383	0.395
8	0.380	0.397	22	0.393	0.399
9	0.369	0.391	23	0.397	0.401
10	0.371	0.385	24	0.406	0.405
11	0.371	0.389	25	0.397	0.399
12	0.369	0.395	26	0.391	0.397
13	0.368	0.383	27	**0.408**	0.407

Table 41: All analyzed Qwen layers for the full-layer BoolQ raw-state linear probe sweep. Cells report test AUC / test balanced accuracy.

L	AUC	Bal. acc.	L	AUC	Bal. acc.
0	**0.424**	0.444	16	0.363	0.389
1	0.407	0.406	17	0.357	0.397
2	0.395	0.419	18	0.366	0.417
3	0.420	0.439	19	0.361	0.397
4	0.409	0.422	20	0.377	0.408
5	0.406	0.444	21	0.370	0.403
6	0.379	0.417	22	0.374	0.414
7	0.375	0.406	23	0.369	0.389
8	0.386	0.414	24	0.370	0.392
9	0.398	0.433	25	0.369	0.392
10	0.380	0.414	26	0.370	0.394
11	0.397	0.433	27	0.363	0.392
12	0.405	0.422	28	0.369	0.397
13	0.396	0.436	29	0.360	0.394
14	0.378	0.403	30	0.363	0.408
15	0.368	0.394	31	0.366	0.411

Table 42: All analyzed Llama layers for the full-layer BoolQ raw-state linear probe sweep. Cells report test AUC / test balanced accuracy.

L	AUC	Bal. acc.	L	AUC	Bal. acc.
0	**0.610**	0.575	14	0.545	0.614
1	0.603	0.578	15	0.539	0.611
2	0.594	0.589	16	0.545	0.592
3	0.589	0.625	17	0.551	0.614
4	0.581	0.608	18	0.548	0.597
5	0.561	0.628	19	0.562	0.586
6	0.559	0.606	20	0.558	0.597
7	0.554	0.614	21	0.550	0.606
8	0.568	0.614	22	0.558	0.597
9	0.573	0.622	23	0.556	0.592
10	0.569	0.614	24	0.556	0.608
11	0.552	0.619	25	0.561	0.628
12	0.561	0.622	26	0.558	0.622
13	0.568	0.614	27	0.561	0.606

Table 43: All analyzed Qwen layers for the full-depth isolated-MLP surrogate-to-differential calibration. Cells report exact-JVP cosine for the single-partition JVP-diagonal fit, centered-IO-diagonal fit, and shuffled-JVP null.

L	JVP	C-IO	Shuffle	L	JVP	C-IO	Shuffle
0	0.185	0.002	0.012	16	0.417	0.406	0.018
1	0.159	0.013	-0.001	17	0.415	0.392	-0.012
2	0.247	0.003	0.024	18	0.398	0.245	0.010
3	0.233	0.034	0.010	19	0.435	0.395	0.036
4	0.275	0.238	0.008	20	0.334	0.263	0.035
5	0.423	0.391	0.028	21	0.243	0.170	0.004
6	0.498	0.434	-0.011	22	0.290	0.098	0.006
7	**0.554**	**0.545**	0.038	23	0.222	0.125	-0.003
8	0.512	0.499	0.077	24	0.154	0.082	0.008
9	0.469	0.457	0.015	25	0.131	0.034	0.004
10	0.461	0.442	-0.002	26	0.122	0.030	-0.003
11	0.491	0.484	0.008	27	0.027	0.015	0.003
12	0.475	0.467	-0.009	28	0.008	0.013	0.000
13	0.454	0.447	**0.105**	29	0.032	0.036	0.001
14	0.384	0.370	0.028	30	0.030	0.037	0.000
15	0.409	0.401	0.028	31	0.083	0.046	0.013

Table 44: All analyzed Llama layers for the full-depth isolated-MLP surrogate-to-differential calibration. Cells report exact-JVP cosine for the single-partition JVP-diagonal fit, centered-IO-diagonal fit, and shuffled-JVP null.

L	JVP	C-IO	Shuffle	L	JVP	C-IO	Shuffle
0	0.066	0.003	-0.000	14	**0.522**	**0.495**	0.026
1	0.149	-0.002	0.005	15	0.452	0.415	0.018
2	0.185	0.028	0.006	16	0.358	0.306	0.026
3	0.259	0.115	0.007	17	0.310	0.257	0.021
4	0.436	0.342	0.003	18	0.246	0.197	0.023
5	0.442	0.234	0.003	19	0.235	0.143	0.005
6	0.464	0.206	0.038	20	0.135	0.033	-0.005
7	0.455	0.444	**0.084**	21	0.063	0.003	-0.006
8	0.459	0.405	0.009	22	0.028	-0.003	0.000
9	0.463	0.453	0.006	23	0.005	-0.003	-0.001
10	0.453	0.447	-0.002	24	0.004	0.005	-0.001
11	0.465	0.456	0.023	25	0.013	0.017	-0.001
12	0.488	0.474	0.014	26	0.049	0.041	-0.002
13	0.505	0.482	0.060	27	0.091	0.045	0.006

Table 45: All analyzed Phi-3 layers for the full-depth isolated-MLP surrogate-to-differential calibration. Cells report exact-JVP cosine for the single-partition JVP-diagonal fit, centered-IO-diagonal fit, and shuffled-JVP null.

L	JVP	C-IO	Shuffle	L	JVP	C-IO	Shuffle
0	0.042	-0.031	-0.004	16	0.263	0.172	0.006
1	0.029	-0.016	0.001	17	0.239	0.144	0.007
2	0.051	0.001	0.004	18	0.221	0.041	0.006
3	0.064	-0.032	0.002	19	0.204	0.007	0.004
4	0.085	-0.001	0.001	20	0.150	-0.045	0.014
5	0.099	-0.022	-0.004	21	0.121	-0.072	0.001
6	0.159	0.019	0.009	22	0.079	-0.061	-0.001
7	0.186	0.009	-0.001	23	0.043	-0.043	-0.002
8	0.230	0.077	0.003	24	0.024	-0.028	-0.001
9	0.243	0.071	0.006	25	0.015	-0.001	0.000
10	0.279	0.187	-0.003	26	0.013	-0.000	-0.000
11	**0.285**	0.228	**0.025**	27	0.010	0.003	0.000
12	0.267	0.218	0.008	28	0.008	0.006	0.000
13	0.259	0.144	0.013	29	0.007	-0.009	0.000
14	0.275	**0.234**	0.010	30	0.000	0.004	-0.000
15	0.271	0.218	0.011	31	0.010	0.011	0.001