Stable Directions, Different Effects: Intervention Fidelity Across Model Precision

Abstract

Activation interventions can be characterized on one numerical realization of a language model and later applied to a quantized version. Preserving standard model quality or direction-level representation similarity does not guarantee that the intervention effect is preserved. We formalize this as cross-precision intervention fidelity and define a signed intervention-fidelity gap between native and transferred steering effects. Across several language models, precision changes, model sizes, steering estimators, and prospectively held-out concepts/tasks, we compare a prespecified direction-alignment/simple-geometry baseline with local functional sensitivity. Intervention effects can change even when directions remain strongly aligned, whereas a standard first-order local approximation tracks the signed gap closely in the strongest held-out evaluations. The novelty is not the Taylor expansion itself, but the intervention-fidelity problem and the empirical accuracy of this local account across the tested replication axes. A correction frozen before fresh evaluation improves reconstruction of native intervention effects and a prespecified sign-based decision. Finally, a prospective parsimony test favors the compact first-order approximation over the richer predictor. Within the tested normalized-transfer regime, intervention reliability should therefore be evaluated functionally rather than inferred from ordinary quantization quality or direction similarity.

1. Introduction

Activation-level interventions offer a lightweight way to change language-model behavior at inference time without updating the underlying weights. Work on contrastive activation addition [1], inference-time intervention [2], representation fine-tuning [3], instruction-specific steering [4], conditional activation steering [5], and fine-grained steering [6] shows that targeted changes can be produced by modifying internal representations [1–6]. These techniques may serve as control techniques and as probes of model computation, but cross-language transfer and unintended effects have also been studied recently [7,8]. However, the numerical realization used to characterize an intervention may not be the same realization deployed in practice. Large language models are often quantized to lower memory consumption and reduce inference cost [9,10] through weight–activation, rotation-based, post-training, mixed-precision, and low-bit schemes [11–17].

This deployment pattern raises an implicit assumption: if a quantized model preserves perplexity, benchmark accuracy, or downstream performance, does an intervention characterized on the higher-precision model preserve its functional effect? Not necessarily. Representation studies show that internal similarity does not guarantee equivalent functional use [18–20]. Information may also remain decodable without being behaviorally required [21–23], which motivates direct intervention-based analyses of internal computation [24,25]. A precision change may therefore leave a steering direction strongly aligned while changing the local function through which that direction acts. We separate three properties: representation stability, intervention fidelity, and behavioral parity. The experiments directly test the first two; behavioral parity is used only as a conceptual distinction, not as a separately established free-generation endpoint. Throughout this work, intervention effect is operationalized as the change in a controlled next-token A/B logit contrast induced by a layerwise activation intervention.

Much of the current work focuses on related questions separately: Steering research develops, localizes and adapts interventions; Representation analyses focus on encoded versus functionally used information; and quantization work focuses on assessing preservation of model quality and efficiency. These are necessary but not specific indicators of the reliability of an intervention before and after a precision change. We investigate how much an intervention that is defined at one precision of weight serves its function after it is transferred to another.

We call this cross-precision intervention fidelity. For a fixed target precision, we compare the effect of a direction characterized and applied natively with the effect of the corresponding direction transferred from another precision. Their signed difference is the intervention-fidelity gap. The sign matters because a precision change can reduce, preserve, or increase the intervention effect, which an unsigned discrepancy would hide.

Across the tested models, model sizes, precision changes, steering estimators, and behavioral concepts, including prospectively held-out concepts and fresh outcomes, steering effects can drift even when directions remain well aligned. The prespecified simple-geometry baseline (cross-precision direction similarity and normalized layer depth, concept, steering estimator, and target precision) is a weak predictor in the strongest held-out evaluations, while the local functional sensitivity obtained from a calibration is much more informative. The same relationship appears across the tested models and the evaluated 1B–3B size range and remains detectable under a distinct LLM.int8() precision change [10], although the gain is weaker there. A correction frozen before fresh evaluation improves native-effect recovery and the prespecified sign decision. A final prospective test of parsimony then shows that the compact first-order functional quantity slightly outperforms the richer predictor.

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

Our contributions are as follows: (1) a direct formulation and measurement of cross-precision intervention fidelity; (2) a conceptual separation of representation stability, intervention fidelity, and behavioral parity, with the empirical tests focused on the first two; (3) evidence that signed fidelity gaps are better accounted for by local functional sensitivity than the prespecified direction-alignment/simple-geometry baseline within the tested FP16↔NF4 regime and an independent FP16↔LLM.int8() condition; and (4) prospective validation followed by a parsimony test, revealing that a compact first-order approximation best captures the most useful transferable signal.

2. Related Work

2.1 Activation Steering and Representation Interventions

Activation steering modifies internal activations at inference time without updating the model weights. Simple activation-space interventions [1,2] have been extended through ReFT, instruction-specific steering, CAST, gradient-guided steering, and selective/fine-grained methods [3–6,26,27]. More recent work studies geometry-aware editing, activation scaling, automated direction discovery, adaptive backtracking, multi-attribute control, and learning-dynamics-based steering [28–34]. Transfer and safety audits also show that steering can generalize while still producing unintended effects [7,8]. This literature mainly asks how to construct, localize, or adapt interventions; by itself, it does not establish whether a fixed intervention keeps the same effect after model precision changes.

2.2 Representation Similarity, Causal Use, and Functional Equivalence

Representation-similarity measures such as CKA describe correspondence between learned spaces [18], but representational correspondence does not necessarily imply the same computational use. Amnesic probing, INLP, and LEACE intervene on encoded information to distinguish decodability from behavioral relevance [21–23]. Causal localization/model editing and mediation analyses similarly target internal computations [24,25]. Recent work also questions cosine similarity as a proxy for functional layer relevance and studies when mechanistic interpretations transfer across networks [19,20]. These findings motivate our separation of representation stability, intervention fidelity, and behavioral parity.

2.3 Quantization and Preservation of Model Function

LLM quantization covers mixed-precision inference and increasingly aggressive low-bit deployment. LLM.int8() and QLoRA are influential mixed-/low-bit baselines [9,10], while GPTQ, AWQ, OmniQuant, SmoothQuant, AQLM, QuaRot, and SpinQuant address weight quantization, activation outliers, calibration, additive compression, and rotations [11–17]. More recent work explores dual transformations, asymmetric calibration, learned flattening, salience-driven mixed precision, sensitivity-aware vector quantization, fractional-bit allocation, and low-precision training [35–41]. Most of these methods are evaluated mainly by model-quality or efficiency criteria. Concurrent work studies activation steering under weight-only INT8/NF4 quantization in 7–9B models and reports highly collinear cross-precision steering vectors together with changed dose-response/capability behavior [42]. Our question is different: we compare target-native and transferred steering within the same target precision and model their signed fidelity gap.

2.4 Novelty Boundary

We do not claim novelty for activation addition, PCA, Fisher/LDA, covariance shrinkage, gradients, Taylor approximations, layer selection, steering, or quantization. Our contribution lies at their intersection: cross-precision intervention fidelity. Rather than using direction alignment or conventional benchmark preservation as a surrogate, we directly compare native and transferred intervention effects and study their signed gap. Our geometry claim is limited to the prespecified simple-geometry baseline; we do not claim to rule out every possible representation-geometric measure.

3. Cross-Precision Intervention Fidelity

3.1 Problem Formulation

Our formulation treats representational correspondence and functional equivalence as separate, consistent with recent functional-similarity analyses [43]. Let p denote the precision used to extract a steering direction and q the precision of the target model. For input x, let $h\ell^{(q)}(x)$ denote the residual representation at layer ℓ, and let $\hat{v}\ell^{(p)}$ denote the unit-norm steering direction extracted at precision p. We intervene at the final prompt token using:

$$h'\ell^{(q)}(x) = h\ell^{(q)}(x) + \alpha \, \|h\ell^{(q)}(x)\|_2 \, \hat{v}\ell^{(p)}, \; \alpha = 0.10.$$

The primary object is unit-normalized steering-direction transfer under target-relative amplitude scaling, since $\hat{v}$ is unit-normalized and the amplitude of the intervention is scaled by the target representation norm; the transfer of the raw source-vector magnitude falls outside the tested estimand.

The controlled one-token A/B score for target option y+ and its opposite y− is

$$s_q(x) = z_q(y+ \mid x) - z_q(y- \mid x).$$

Here z_q is the next-token logit; because both candidates share the same next-token normalization, this logit difference is also their log-probability difference. In the next equation, s_q(h) denotes the downstream score when the layer-ℓ residual state is set to h. The item-level and mean steering effects are:

$$\tau\ell,x^{(p\to q)} = s_q(h\ell^{(q)} + \alpha \|h\ell^{(q)}\|_2 \hat{v}\ell^{(p)}) - s_q(h\ell^{(q)}),\ T\ell^{(p\to q)} = E_x[\tau\ell,x^{(p\to q)}].$$

We use p and q for generic extraction and target precisions; for a particular transfer comparison, s denotes the source precision and t the target precision.

For target precision t and alternative source precision $s \neq t$, the primary estimand is the signed intervention-fidelity gap

$$D\ell^{(t)} = T\ell^{(t\to t)} - T\ell^{(s\to t)}.$$

Thus D=0 denotes exact native/transferred equality; under the fixed score orientation, the sign records whether transfer numerically reduces or increases the intervention effect relative to native. Absolute |D| is secondary. The distinction and reconstruction logic is summarized in Figure 1.

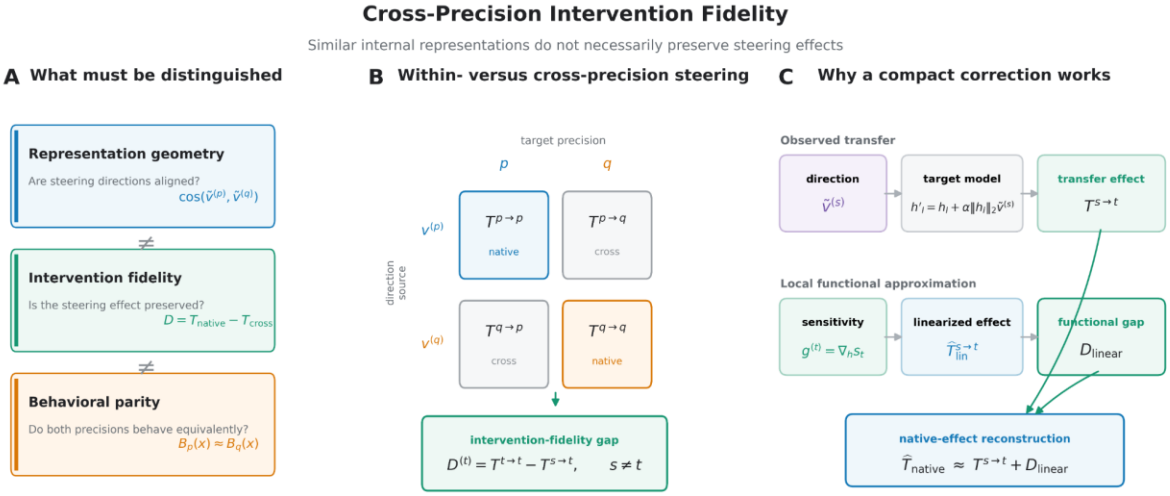

Figure 1. Conceptual overview. (A) There is a conceptual distinction between representation stability, intervention fidelity, and free-generation behavioral parity; representation/direction stability and intervention fidelity are directly tested, whereas free-generation behavioral parity is not directly tested. Signed gap D^(t) is defined by (B) native and cross-precision effects. D_linear with native-effect reconstruction is obtained from functional sensitivity at the local level (C).

3.2 Steering Estimation and Functional Sensitivity

We use three established linear estimators as controlled instruments: contrastive activation addition (CAA) [1], pair-difference Δ-PCA, and shrinkage Fisher/LDA with Oracle Approximating Shrinkage [44]. All directions are L_2-normalized; the full estimator equations are given in Appendix A.

The prespecified direction-alignment/simple-geometry baseline is weakly predictive of the signed fidelity gap in the strongest held-out evaluations, motivating a complementary functional description. Let $g\ell, x^{(q)}=\nabla_\{h\ell^{(q)}\}\ s_q(x)$ denote the local sensitivity of the target score to the intervened residual state. The relation below is the standard first-order Taylor approximation, not a new mathematical construction. Our empirical question is whether this local approximation remains quantitatively informative across the numerical model realizations, steering estimators, layers, and semantic domains evaluated here:

$$\tau\ell,\ x^{(p\to q)} \approx \alpha \|h\ell^{(q)}(x)\|_2\ (g\ell,\ x^{(q)})^{\top}\ \hat{v}\ell^{(p)}.$$

Averaging over calibration items gives $\tilde{T}\ell^{(p\to q)}$ and

$$D_linear,\ \ell^{(t)} = \tilde{T}\ell^{(t\to t)} - \tilde{T}\ell^{(s\to t)}.$$

This motivates the reconstruction $T_native = T^{(s\to t)} + \hat{D}$. Because D_linear combines target-model sensitivity with source and target-native directions, it is a diagnostic/effect-reconstruction procedure rather than a replacement for target-side direction extraction. **C**omputing D_linear requires target-model gradients on calibration examples and therefore incurs backward-pass and calibration cost; the predictor-stage timings reported later exclude this upstream computation. Before external replication, the predictor feature families and evaluation protocol were frozen; this does not imply that one unchanged coefficient vector was carried through every replication stage. The later fresh recovery and parsimony analyses provide the stronger prediction-frozen tests (Appendices A and F).

3.3 Models, Precision Changes, and Generalization Axes

Mechanism discovery uses Qwen2.5-1.5B-Instruct [45] under FP16↔NF4 across all 28 transformer layers. Independent model replication uses Llama-3.2-1B-Instruct [46] across all 16 layers, whereas model-size and prospective concept/task evaluations use Llama-3.2-3B-Instruct across all 28 layers. NF4 is the 4-bit NormalFloat datatype introduced with QLoRA [9]; here it is weight-only quantization with FP16 computation. FP16 is the reference precision, not a claim of BF16 equivalence. A quantizer dependence test is also performed separately in FP16↔LLM.int8() [10]. Age/merit, refusal/safety, and sentiment are initial concepts, while the prospective semantic holdout employs factual truthfulness, instruction compliance, and uncertainty calibration. This gives 3×3×28×2 = 504 structured concept–estimator–layer–target cells, which are not treated as 504 IID experiments for inferential uncertainty. Table 1 shows the frozen sequence.

Table 1. Frozen experimental scope. All steering uses CAA, Δ-PCA, and shrinkage Fisher with α=0.10. NF4 is weight-only with FP16 computation.

Study	Model / precision	Concepts / tasks	Layers	Evaluation / endpoint
Discovery	Qwen2.5-1.5B, FP16↔NF4	age/merit; refusal/safety; sentiment	28	held-out layer-CV; signed D
Cross-model	Llama-3.2-1B, FP16↔NF4	same initial concepts	16	held-out layer-CV; signed D
Cross-quantization	Llama-3.2-1B, FP16↔INT8	same initial concepts	16	held-out layer-CV; signed D
Model-size (1B→3B)	Llama-3.2-3B, FP16↔NF4	same initial concepts	28	held-out layer-CV; signed D
Held-out concepts	Llama-3.2-3B, FP16↔NF4	truthfulness; compliance; uncertainty	28	prospective holdout; signed D
Recovery / decision / parsimony	Llama-3.2-3B, FP16↔NF4	held-out concepts above	28	fresh items; RMSE / sign errors

3.4 Anti-Leakage Design and Inference

Functional predictors are built from data that are disjoint from the outcomes they predict. In discovery, each concept uses 32 calibration items and a separate 32-item prediction set. The calibration set estimates calibration-averaged functional quantities rather than serving as an inferential sample in its own right; uncertainty is evaluated through held-out prediction and layer-cluster resampling rather than by treating calibration items or resulting cells as IID observations. Evaluation uses seven-fold grouped cross-validation by layer, so repeated cells from the same layer cannot cross training and validation folds. The same separation is kept for external replication. Before external replication, the feature families/model class, α, three estimators, signed D, anti-leakage rule, and evaluation criteria were frozen. Those model/quantizer/size/task stages test replication of the prespecified predictive relationship; they are not presented as coefficient-free transfer of one immutable fit. The later recovery test evaluates an immutable correction on fresh outcomes without predictor refitting, and the final parsimony test freezes all comparator predictions before fresh outcomes (Appendix F). No primary analysis removes near-zero cells post hoc, selects favorable layers, drops an estimator because of its results, or replaces signed D with |D|.

The main comparisons use held-out R^2, RMSE, $\Delta R^2 = R^2_{aug} - R^2_{geom}$, and relative RMSE reduction. Uncertainty is estimated with layer-level or layer-cluster bootstrap resampling rather than by treating repeated cells as IID. In the 504-cell held-out/fresh analyses, the cells are structured repetitions across layers, concepts, estimators, and target precisions. For the decision consequence, a 10,000-draw layer-cluster bootstrap is the primary dependence-aware uncertainty summary, while exact McNemar inference is a complementary paired-cell diagnostic. The fresh parsimony test uses a 10,000-draw layer-cluster bootstrap for the paired compact-minus-full RMSE difference. Extended definitions are given in Appendix B.

4. Results

Table 2 summarizes the key held-out and prospective evidence. The signed intervention-fidelity gap is denoted by D, and "drift" is used descriptively here to denote variations across precision.

Table 2. Decisive evidence is the value of functional sensitivity over and above the prespecified simple-geometry baseline (ΔR^2). The final row preserves the prospective negative result.

Test	Decisive statistic	Uncertainty / test	Supported conclusion
Discovery	geom R^2=.010; aug R^2=.926; ΔR^2=.916	ΔR^2 95% CI [.870,.947]; RMSE -72.61%	functional sensitivity explains most signed gap
Cross-model	geom -.007; aug .864; ΔR^2=.870	ΔR^2 95% CI [.767,.928]; RMSE -63.21%	replicates in independent model
INT8	ΔR^2=.340; grad cosine .971	ΔR^2 95% CI [.118,.528]; RMSE -18.43%	weaker but positive cross-quantizer evidence
Model-size	geom .007; aug .989; ΔR^2=.982	ΔR^2 95% CI [.969,.996]; RMSE -89.32%	persists from 1B to 3B in tested family
Held-out concepts	geom .001; aug .966; ΔR^2=.965; r(D linear, D) =.984	ΔR^2 95% CI [.946,.974]; RMSE -81.48%	generalizes to held-out concepts/tasks
Recovery	RMSE .0667→.0264; corrected R^2=.975	RMSE -60.42%; CI [51.98,66.83] %	frozen correction recovers native effects
Decision	wrong 12.50%→6.75%; -5.75 pp	layer-bootstrap CI [1.98,9.72] pp; P(reduction>0) =.9982; McNemar p=.0009143 (secondary)	wrong-decision rate reduced 46.03%
Parsimony	compact .0267; full .0269; Δ=-.000174	10k layer-bootstrap CI [-.000314, -.000056]; tail P(Δ>0) =.0013	richer predictor does not win

4.1 Steering Can Drift Despite Direction-Level Stability

Signed gaps vary systematically across layers, steering estimators, concepts, and target precisions rather than collapsing around zero (Figure 2A).This variation is not specified by the prespecified direction-alignment/simple-geometry descriptors and can also have significant gaps despite a strong cross-precision direction alignment (Figure 2B). The independent FP16↔LLM.int8() condition is a useful boundary case: mean functional-gradient cosine is 0.971, yet effects are not preserved exactly; functional information still adds ΔR^2=0.340 (95% CI [0.118,0.528]) and reduces prediction RMSE by 18.43% (Figure 2C; Table 2). Strong direction alignment therefore does not by itself establish intervention fidelity. This claim is restricted to the evaluated precision transfers and does not imply that quantization generally disrupts activation steering.

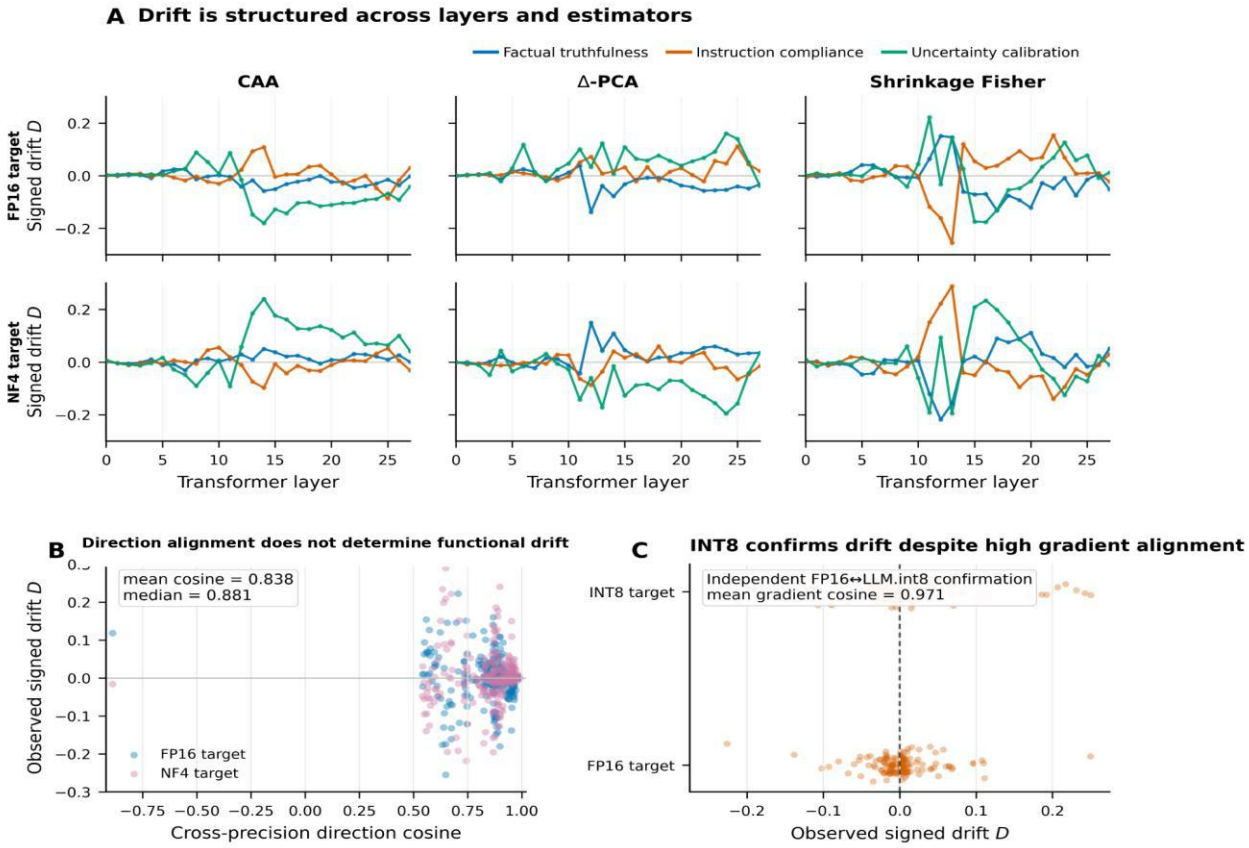

Figure 2. Structured cross-precision gap by direction alignment (but not certified). The values of (A) Signed D vary from layer to estimator. D can coexist with (B) Strong direction alignment. (C) The independent INT8 condition can be maintained with high mean gradient alignment and there is still residual gap.

4.2 Functional Sensitivity Outpredicts the Prespecified Simple-Geometry Baseline

In discovery, the direction-alignment/simple-geometry baseline accounts for a very small amount of the held-out variation (R^2=.010), while the baseline plus calibration-derived functional sensitivity increases R^2 to .926 (ΔR^2=+.916; 95% CI [.870,.947]) and decreases RMSE by 72.61%. The contrast is strongest on the 504 prospectively held-out concept/task cells: simple-geometry R^2=.001, augmented R^2=.966, and the compact first-order quantity itself closely tracks observed D, with r(D_linear, D)=.984 (Figure 3A–B). Functional sensitivity therefore supplies nearly all the predictive information missing from the prespecified simple-geometry baseline in these held-out tests; this does not mean that every possible representation-geometric metric is non-predictive.

4.3 The Functional Relation Replicates Across Evaluation Axes

The relationship persists across changes in model, quantizer, model size within the tested 1B–3B range, and semantic task, although its strength is not invariant (Figure 3C–D; Table 2). Independent Llama-3.2-1B FP16/NF4 replication gives ΔR^2=.870 and a 63.21% RMSE reduction; the distinct INT8 condition is weaker (ΔR^2=.340, 18.43% reduction); Llama-3.2-3B model-size evaluation strengthens the relation (ΔR^2=.982, 89.32% reduction); and prospectively held-out truthfulness, compliance, and uncertainty cells give ΔR^2=.965, 81.48% RMSE reduction, and r (D_linear, D)=.984. These results support generalization across the tested axes without implying quantizer-independent effect size.

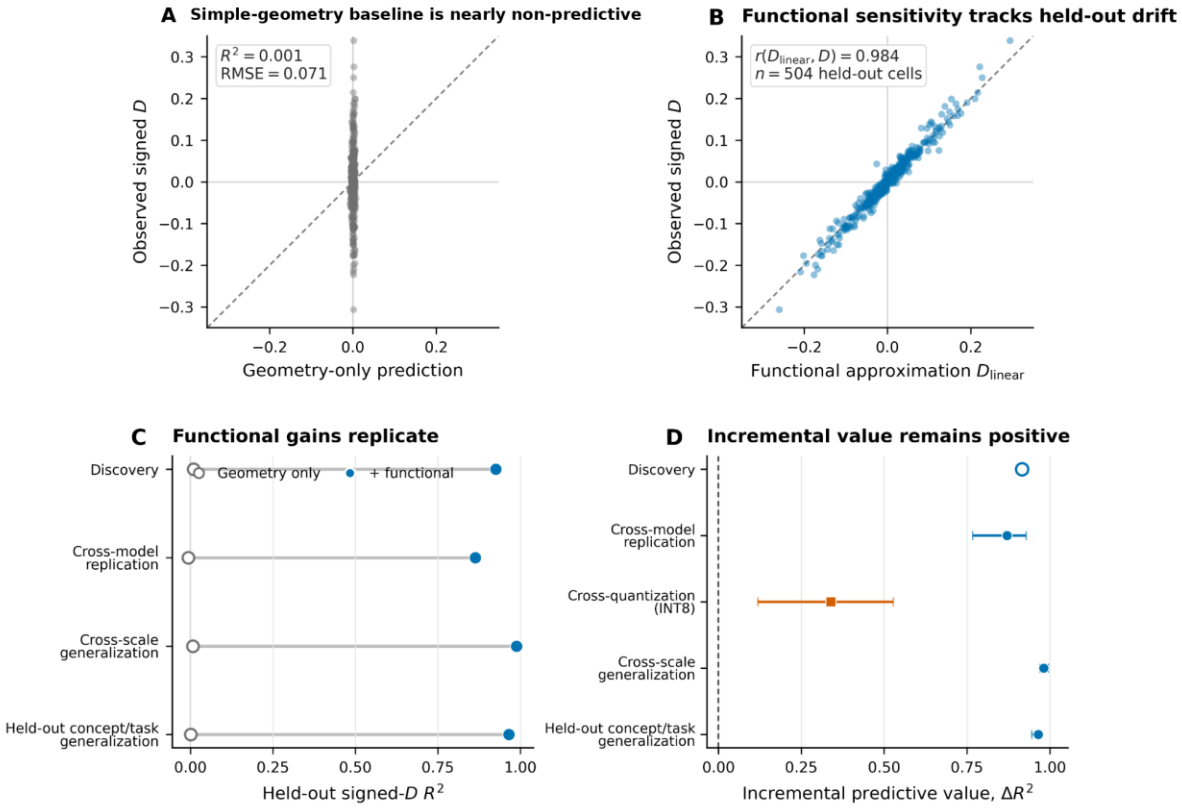

Figure 3. Functional sensitivity tracks held-out signed gaps across tested axes. (A) The prespecified simple-geometry prediction is nearly non-predictive on held-out concepts/tasks. (B) D_linear closely tracks observed D. (C–D) Functional gains remain positive across discovery and replication axes, with the weaker INT8 case retained.

4.4 Frozen Correction Improves Prospective Recovery and Decisions

We freeze the correction before evaluating fresh outcomes. Uncorrected native-effect recovery has RMSE 0.0667; the frozen correction reduces it to 0.0264, a 60.42% reduction (95% CI [51.98,66.83] %), and reaches R^2=.975 against the observed native effects (Figure 4A–C). Under the prospectively frozen zero-threshold sign rule, wrong decisions fall from 12.50% to 6.75%, an absolute reduction of 5.75 percentage points and a relative reduction of 46.03%. The primary dependence-aware summary is the 10,000-draw layer-cluster bootstrap: 95% CI [1.98,9.72] percentage points and P(reduction>0) =.9982. Exact McNemar p=.0009143 is reported only as a complementary paired-cell diagnostic (Figure 4D). This is an operational result for one prespecified rule, not a general deployment-safety guarantee.

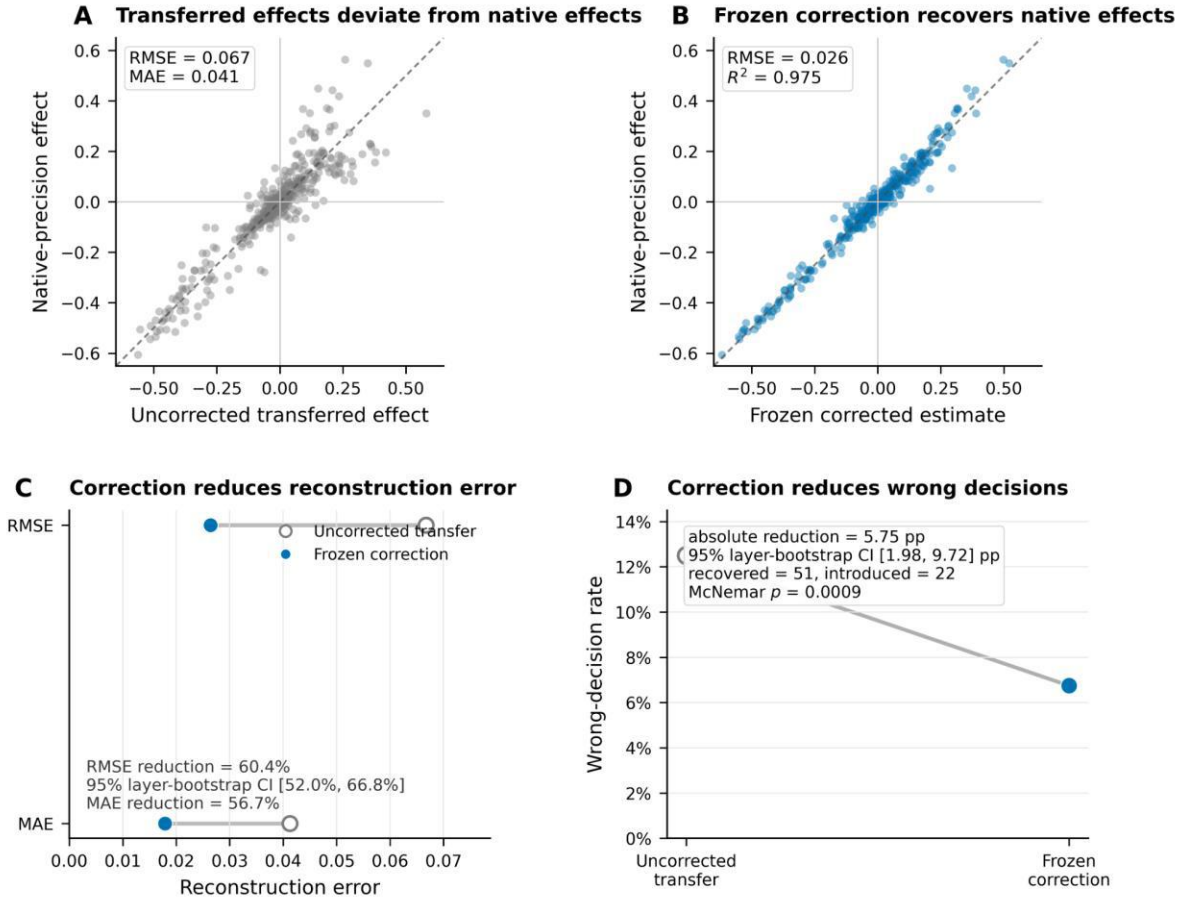

Figure 4. Frozen correction improves fresh native-effect recovery and the prespecified sign decision. (A) Transferred effects deviate from native effects. (B–C) Correction reduces reconstruction error. (D) Wrong-decision rate falls under the frozen zero-threshold rule.

4.5 A Fresh Parsimony Test Favors the Compact Approximation

The final comparison asks whether a richer multifeature predictor is necessary. It is not. The one-feature D_linear predictor has fresh-data RMSE 0.0267 versus 0.0269 for the full predictor. The prespecified difference, RMSE_D_linear−RMSE_full=−0.000174, has a 10,000-draw layer-cluster-bootstrap 95% CI [−0.000314, −0.000056]; P(Δ>0) =0.0013 is the bootstrap tail probability that compact-minus-full RMSE is positive (Figure 5A–B). The frozen superiority hypothesis thus fails in the opposite direction in the case of the richer predictor. This negative result makes the account much simpler because much of the useful transferable signal is picked up by local first order sensitivity and not by multifeature complexity. D_linear is using one fitted coefficient instead of 18 in the full model; predictor-stage time is about 2.63 versus 5.80 ms for 504 predictions (excluding upstream gradient, activation, calibration, direction-construction and intervention costs).

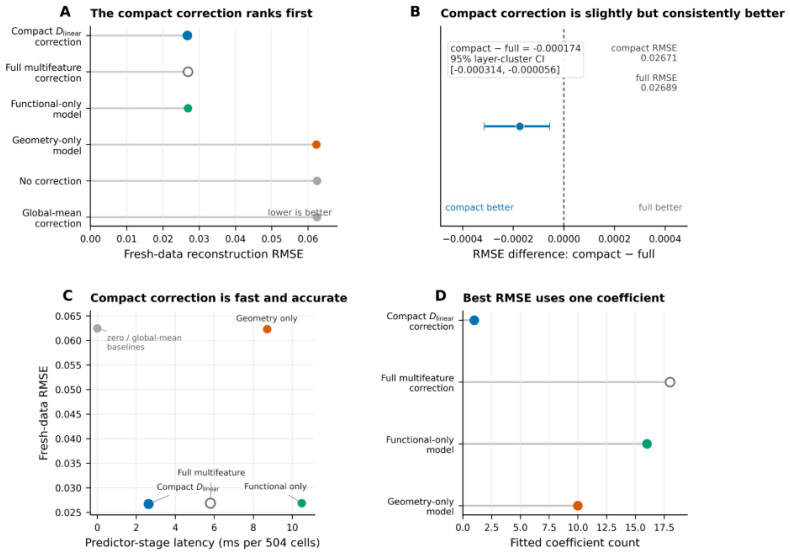

Figure 5. Prospective parsimony favors compact correction. (A) D_linear performs best in terms of fresh recovery RMSE. (B) Its paired RMSE advantage over the full model is small but consistently below zero. At the predictor stage, the compact predictor is simple and fast (C–D).

5. Discussion and Limitations

A steering direction can remain strongly aligned across precision while the effect produced along that direction changes; high direction alignment therefore does not certify intervention fidelity in the evaluated transfers. This further operationalizes the distinction between representational correspondence and functional equivalence [19,20]. The independent INT8 result is an informative boundary observation: mean gradient alignment remains high, yet fidelity is not exact and the remaining signed gap is still predictably related to functional information. The weaker gain may reflect structural differences in LLM.int8()'s outlier-aware mixed-precision computation, but the present study does not isolate that mechanism; we therefore treat it as evidence that predictive strength is quantizer-dependent rather than as a causal explanation. Our question is distinct from demonstrations of behavioral or linguistic steering transfer [7]: transfer can succeed while its magnitude changes enough to require explicit fidelity evaluation.

The strong performance of D_linear should be interpreted as an empirical result about cross-precision intervention fidelity, not as novelty in gradients or Taylor approximation. Whether the first order expression is still quantitatively meaningful after model precision changes, model family changes, model size changes, steering estimator changes and changes in the semantic domain is an empirical question that arises from the standard local analysis. The neglected Taylor remainder is a function of the size of the perturbation, as well as local curvature, so that the quality of the approximation does not necessarily have to hold as the steering strength is increased significantly, where higher order or nonlinear fidelity models may be needed. The prospective held-out and fresh-outcome evaluations therefore provide stronger evidence for the claim than the derivation itself. The final negative parsimony result further narrows the interpretation: If geometric, structural and further functional features are added, no better prediction of the fresh data is achieved than with the compact approximation. This does not guarantee that the general case of D_linear is optimal.

The main sequence employs FP16/NF4, discovery by Qwen2.5, "replication" by Llama-3.2, model-size evaluation: only 1B-3B, and 1 FP16↔LLM.int8() confirmation. Broader weight–activation and rotation-based regimes remain untested [11,12] and a BF16-equivalence claim is not made. The endpoint is an A/B logit contrast (one token) with one frozen strength, α=.10, at the final prompt-token representation. Normalized transfer does not involve raw-magnitude transfer as it is target-relative amplitude scaled. The present study does not provide unconstrained generation behavioral parity, leaving open a trajectory-level question: local fidelity gaps may compound autoregressively. The results do not establish behavior at other token positions or under nonlinear/adaptive steering [27]. The zero-threshold decision is one of the pre-specified rules and not an overall safety guarantee. Layer-cluster inference is performed at the layer level (where repeated structure is considered) with generalization to the models, quantizers, estimators, and behavioral domains remaining empirical. D_linear's upstream diagnostic cost is required for target side gradient computation and is not the end-to-end system cost reported with the predictor side timing. Further research is needed to describe fidelity as a function of intervention strength, to examine unconstrained autoregressive generation and to explore the analysis with respect to wider ranges of quantization regimes and more highly structured steering directions. The extended scope boundaries are given in Appendix E.

6. Conclusion

We question whether an activation-steering intervention specified at one model precision has the same functional impact when applied to other numerical models. Fidelity is not necessarily related to direction similarity and/or the prespecified simple-geometry baseline across the tested models, 1B–3B size range, estimators, precision changes, and held-out concepts: similar directions can yield different intervention effects. Local functional sensitivity provides substantially more information about the signed intervention-fidelity gap. A correction performed before re-evaluation of the native effects yields better recovery of the native effects and the prespecified sign decision, while a separate prospective parsimony test favors the simpler first-order predictor over the more complicated model. These results justify considering transferred activation interventions based on their functional impact, not only on the direction alignment or on traditional quantization quality, within the tested normalized-transfer regime.

Reproducibility Statement

The definition of the intervention, the estimates of the steering indicators, the settings for precision, calibration/outcome separation and grouped evaluation, statistical procedures and prospective freezing rules are specified in Sections 3-4, Tables 1-2 and Appendices A-D and F. Changes in model, quantizer, model size, and concept/task domain are treated as replication axes. Appendix F distinguishes protocol-frozen from prediction-frozen stages; the fresh native-effect correction and final compact-versus-full predictions are fixed before their corresponding fresh outcomes are evaluated. No additional code-release commitment is made beyond the materials included in this submission.

Ethics / Broader Impact Statement

Activations may affect the behavior of a model in ways that are either undesirable or desirable, and numerical alterations may affect that behavior even if there are no corresponding changes in the usual measures of model quality. The study thus provides backing for revalidation when interventions are rolled out into compressed deployments. It is not a general deployment safety rule, and the correction will not prevent harmful activity, because its controlled A/B endpoints and single sign-based decision rule are not general safety rules. The contribution is diagnostic: it reveals whether a numerical realization of an intervention acts as a faithful approximation of another.

AI Use Statement

Generative AI tools were used only to enhance the language, readability, and clarity of the manuscript.

References

[1] N. Rimsky, N. Gabrieli, J. Schulz, M. Tong, E. Hubinger, and A. Turner. Steering Llama 2 via Contrastive Activation Addition. ACL, pp. 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828.

[2] K. Li, O. Patel, F. Vi'egas, H. Pfister, and M. Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. NeurIPS 36, pp. 41451–41530, 2023.

[3] Z. Wu, A. Arora, Z. Wang, A. Geiger, D. Jurafsky, C. D. Manning, and C. Potts. ReFT: Representation Finetuning for Language Models. NeurIPS 37, 2024.

[4] A. Stolfo, V. Balachandran, S. Yousefi, E. Horvitz, and B. Nushi. Improving Instruction-Following in Language Models through Activation Steering. ICLR, 2025.

[5] B. W. Lee, I. Padhi, K. N. Ramamurthy, E. Miehling, P. Dognin, M. Nagireddy, and A. Dhurandhar. Programming Refusal with Conditional Activation Steering. ICLR, 2025.

[6] Z. Feng et al. Fine-Grained Activation Steering: Steering Less, Achieving More. ICLR, 2026.

[7] X. Wang, M. Wang, Y. Liu, H. Schuetze, and B. Plank. Refusal Direction is Universal Across Safety-Aligned Languages. NeurIPS 38, 2025.

[8] Y. Li, A. Fastowski, E. Zaradoukas, B. Prenkaj, and G. Kasneci. Analysing the Safety Pitfalls of Steering Vectors. Findings of ACL, pp. 11182–11204, 2026. doi: 10.18653/v1/2026.findings-acl.544.

[9] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs. NeurIPS 36, 2023.

[10] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer. LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. NeurIPS 35, 2022.

[11] G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, and S. Han. SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models. ICML, PMLR 202:38087–38099, 2023.

[12] S. Ashkboos et al. QuaRot: Outlier-Free 4-Bit Inference in Rotated LLMs. NeurIPS 37, 2024.

[13] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. ICLR, 2023.

[14] J. Lin et al. AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration. MLSys, vol. 6, 2024.

[15] W. Shao et al. OmniQuant: Omnidirectionally Calibrated Quantization for Large Language Models. ICLR, 2024.

[16] V. Egiazarian, A. Panferov, D. Kuznedelev, E. Frantar, A. Babenko, and D. Alistarh. Extreme Compression of Large Language Models via Additive Quantization. ICML, PMLR 235:12284–12303, 2024.

[17] Z. Liu et al. SpinQuant: LLM Quantization with Learned Rotations. ICLR, 2025.

[18] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of Neural Network Representations Revisited. ICML, PMLR 97:3519–3529, 2019.

[19] C. Hinostroza et al. Rethinking Layer Relevance in Large Language Models Beyond Cosine Similarity. ICLR, 2026.

[20] A. Sun and M. Toneva. Tracking Equivalent Mechanistic Interpretations Across Neural Networks. ICLR, 2026.

[21] Y. Elazar, S. Ravfogel, A. Jacovi, and Y. Goldberg. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. TACL, 9:160–175, 2021. doi:10.1162/tacl_a_00359.
[22] S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. ACL, pp. 7237–7256, 2020.
[23] N. Belrose, D. Schneider-Joseph, S. Ravfogel, R. Cotterell, E. Raff, and S. Biderman. LEACE: Perfect Linear Concept Erasure in Closed Form. NeurIPS 36, pp. 66044–66063, 2023.
[24] K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and Editing Factual Associations in GPT. NeurIPS 35, pp. 17359–17372, 2022.
[25] A. Mueller et al. The Quest for the Right Mediator: Surveying Mechanistic Interpretability for NLP Through the Lens of Causal Mediation Analysis. Computational Linguistics, 52(1):331–378, 2026.
[26] D. Nguyen, A. Prasad, E. Stengel-Eskin, and M. Bansal. GrAInS: Gradient-based Attribution for Inference-Time Steering of LLMs and VLMs. ACL, pp. 46523–46543, 2026.
[27] Q.-A. Dang and C. Ngo. Selective Steering: Norm-Preserving Control Through Discriminative Layer Selection. Findings of ACL, pp. 10887–10910, 2026.
[28] V.-C. Pham and T. H. Nguyen. Householder Pseudo-Rotation: A Novel Approach to Activation Editing in LLMs with Direction-Magnitude Perspective. EMNLP, pp. 13737–13751, 2024.
[29] N. Stoehr, K. Du, V. Snæbjarnarson, R. West, R. Cotterell, and A. Schein. Activation Scaling for Steering and Interpreting Language Models. Findings of EMNLP, pp. 8189–8200, 2024.
[30] V. Siu et al. COSMIC: Generalized Refusal Direction Identification in LLM Activations. Findings of ACL, pp. 25534–25553, 2025.
[31] M. H. Vu and T. M. Nguyen. Angular Steering: Behavior Control via Rotation in Activation Space. NeurIPS 38, 2025.
[32] Z. Cheng et al. Steering When Necessary: Flexible Steering Large Language Models with Backtracking. NeurIPS 38, 2025.
[33] N. F. Oozeer, L. Marks, F. Barez, and A. Abdullah. Beyond Linear Steering: Unified Multi-Attribute Control for Language Models. Findings of EMNLP, pp. 23513–23557, 2025.
[34] K. Sharma and R. Trivedi. COLD-Steer: Steering Large Language Models via In-Context One-step Learning Dynamics. ICLR, 2026.
[35] H. Lin et al. DuQuant: Distributing Outliers via Dual Transformation Makes Stronger Quantized LLMs. NeurIPS 37, 2024.
[36] Y. Li, R. Yin, D. Lee, S. Xiao, and P. Panda. GPTAQ: Efficient Finetuning-Free Quantization for Asymmetric Calibration. ICML, PMLR 267:36690–36706, 2025.
[37] Y. Sun et al. FlatQuant: Flatness Matters for LLM Quantization. ICML, PMLR 267:57587–57613, 2025.
[38] W. Huang et al. SliM-LLM: Salience-Driven Mixed-Precision Quantization for Large Language Models. ICML, PMLR 267:25672–25692, 2025.
[39] Z. Xu, X. Hu, Q. Wu, and D. Yang. RSAVQ: Riemannian Sensitivity-Aware Vector Quantization for Large Language Models. NeurIPS 38, 2025.
[40] D. Lee and H. O. Song. Q-Palette: Fractional-Bit Quantizers Toward Optimal Bit Allocation for Efficient LLM Deployment. NeurIPS 38, 2025.
[41] B. Chmiel, M. Fishman, R. Banner, and D. Soudry. FP4 All the Way: Fully Quantized Training of Large Language Models. NeurIPS 38, 2025.
[42] S. Bhandari and B. Wade. Steering Under Compression: Dose-Response, Capability Cost, and Failure Asymmetry in Quantized LLMs. arXiv:2609.06473, 2026. Concurrent preprint.
[43] I. Athanasiadis, A. Karmush, and M. Felsberg. Grounding Functional Similarity by Invariance-Aware Model Stitching. ICML, 2026.
[44] Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero. Shrinkage Algorithms for MMSE Covariance Estimation. IEEE Transactions on Signal Processing, 58(10):5016–5029, 2010.
[45] A. Yang et al. Qwen2.5 Technical Report. arXiv:2412.15115, 2024. Model provenance source.
[46] Meta AI. Llama 3.2 Model Card. 2024. Official model provenance source.

Appendix A. Detailed Steering Estimators and Functional Predictors

For each concept, steering-direction extraction used 128 frozen contrastive pairs. For pair (x_i^+, x_i^-),

$$\delta_{\ell,i}^{(p)} = h_\ell^{(p)}(x_i^+) - h_\ell^{(p)}(x_i^-), \quad v_\ell, \mathrm{CAA}^{(p)} = (1/N) \Sigma_i \delta_{\ell,i}^{(p)}.$$

Pair-difference PCA uses the leading component of centered $\delta_{\ell,i}^{(p)}$, with sign aligned to CAA. Shrinkage Fisher uses

$$v_\ell, \mathrm{Fisher}^{(p)} \propto \hat{\Sigma}_\ell^{(-1)} (\mu_{\ell,+}^{(p)} - \mu_{\ell,-}^{(p)}).$$

with Oracle Approximating Shrinkage [44]. All directions are normalized by L_2. Calibration-averaged first-order effects are

$$\tilde{T}\ell^{(p\to q)} = E_{x\in C}\,[\alpha\,\|h\ell^{(q)}(x)\|_2\,(g\ell,x^{(q)})^{T}\,\hat{v}\ell^{(p)}].$$

The prespecified simple-geometry baseline includes cross-precision direction similarity together with normalized layer depth, concept, estimator, and target precision. The augmented predictor adds own-/cross-precision functional alignment, their difference, native/transferred linearized effects, D_linear, mean gradient norm, and mean hidden-state norm. Predictor families were frozen before external replication. When a fitted linear correction is used, we distinguish the derived D_linear quantity from the fitted predictor and write the predictive form generically as $\hat{D}=\beta_0+\beta^{T}x$. The compact comparator uses x=(D_linear); the full predictor uses the feature family above. The frozen prospective parsimony artifact contains one learned numeric coefficient for the compact comparator and 18 coefficients for the full predictor. Exact preprocessing, intercept handling, categorical coding, and regularization settings are recorded in the frozen implementation artifacts and were held fixed for the reported analyses.

Appendix B. Extended Evaluation and Statistical Definitions

The naive recovery estimate is T_naive=T^(s→t) and the corrected estimate is T_native=T^(s→t) +$\hat{D}$, both compared with the subsequently observed T^(t→t). Recovery is evaluated with RMSE, MAE, and R^2 using layer-cluster bootstrap uncertainty. The 504-cell analyses consist of structured repetitions over 28 layers, three concepts, three estimators, and two target precisions; they are not treated as IID for inferential uncertainty. The decision analysis applies to a prospectively frozen zero threshold. Its primary uncertainty summary is a 10,000-draw layer-cluster bootstrap of the absolute error-rate reduction, while exact McNemar inference is reported as a complementary paired-cell diagnostic. The fresh parsimony analysis compares the full predictor, one-feature D_linear, functional-only, simple-geometry, global-mean, and zero-gap baselines. Its primary endpoint is the paired fresh-data RMSE difference with a 10,000-draw layer-cluster bootstrap; P(Δ>0) is the fraction of bootstrap draws in which compact-minus-full RMSE is positive. Predictor-stage timing excludes gradient computation, hidden-state collection, direction construction, calibration, and intervention inference.

Appendix C. Reproducibility and Preprocessing Details

Each concept uses 128 frozen contrastive extraction pairs for steering-direction estimation. In mechanism discovery, functional predictors use 32 calibration items and intervention outcomes use a disjoint 32-item prediction set; the same separation principle is retained in later replication and prospective stages. Prompts are tokenized using the checkpoint-specific tokenizer/chat-template convention frozen in the experiment scripts, and activations are read at the final prompt token after that rendering. Candidate-label tokenization was checked under each checkpoint-specific tokenizer and rendered chat template before outcome evaluation so that the controlled endpoint referred to the intended single next-token A/B candidates. The controlled endpoint compares the next-token logits of the two predeclared A/B candidate tokens under the same rendered prompt; no free-generation decoding is used for the primary endpoint. The normalized intervention uses α=.10 at one layer at a time, with all other model parameters frozen. No post-outcome layer selection, estimator dropping, near-zero-cell filtering, or signed-to-absolute endpoint substitution is performed. The six concept/task families are age/merit, refusal/safety, sentiment, factual truthfulness, instruction compliance, and uncertainty calibration. Their prompt construction and source provenance are defined by the frozen prompt-generation/data records used throughout the experiments; we do not reconstruct or modify these definitions post hoc. The exact checkpoint revisions, software/package versions, quantizer arguments, tokenizer/chat-template strings, A/B token IDs, seeds, split manifests, preprocessing code, and predictor-fitting configuration are defined by the frozen implementation records rather than reconstructed in manuscript prose.

Appendix D. Extended Replication and Predictor Complexity

Independent Llama-3.2-1B FP16/NF4 replication yields simple-geometry R^2=-.007, augmented R^2=.864, ΔR^2=.870 (ΔR^2 95% CI [.767,.928]), and 63.21% RMSE reduction. FP16/LLM.int8 **()** yields ΔR^2=.340 (ΔR^2 95% CI [.118,.528]), 18.43% RMSE reduction, and mean gradient cosine .971. Llama-3.2-3B model-size evaluation yields simple-geometry R^2=.007, augmented R^2=.989, ΔR^2=.982 (ΔR^2 95% CI [.969,.996]), and 89.32% RMSE reduction. On 504 prospective held-out concept/task cells, simple-geometry R^2=.001, augmented R^2=.966, ΔR^2=.965 (ΔR^2 95% CI [.946,.974]), 81.48% RMSE reduction, and r(D_linear, D) =.984. D_linear uses one fitted coefficient versus 18 for the full model; predictor-stage CPU time is approximately 2.63 versus 5.80 ms for 504 predictions.

Appendix E. Extended Scope Limitations

Modern quantization includes other regimes than those tested here, such as joint weight–activation quantization and rotation-based low-bit inference [11,12]. FP16 is the high precision reference, no BF16-equivalence claim. The empirical geometry comparison is for a given direction-alignment/simple-geometry baseline and not for all possible CKA/subspace/covariance measures. The controlled one-token A/B endpoint does not create the behavioral parity, long-

form reasoning behavior or open-ended decoding of free-generation. Other strengths, such as token positions, raw-magnitude transfer conventions, nonlinear operators, and adaptive layer-selection procedures, may exhibit other behaviors; the later program takes one of these strengths as its basis, α=.10. No report of a dedicated post-freeze steering-strength sweep, empirical-calibation comparator, co-registered conventional-quality parity study, 7B/8B replication, or GPTQ/AWQ/free-generation confirmation are reported, which are scope extensions, not part of the frozen discovery-to-prospective evaluation sequence. This is a single rule for the decision of signing, not a universal deployment-safety rule. The generalization of these models, their sizes 1B-3B, their quantizers, their estimators, and their behavioral domain is still empirical. Predictor-stage timing isn't an end-to-end system-cost measurement.

Appendix F. Prospective Freeze and Fit Ledger

The following table differentiates claims that are protocol-frozen and claims that are prediction-frozen. External replication does not imply that a single coefficient vector was replicated for each model, quantiser, size condition, but rather that the definition of the intervention, the signed endpoint, the estimators, the feature families/model class, the anti-leakage rule and the evaluation logic are replicated. The best immutable-prediction tests are the fresh recovery and parsimony stages.

Stage	**Frozen before outcome**	**Coefficient / prediction status**	**Outcome protection**
Discovery	signed D, estimators, split/evaluation design established	fitted/evaluated by layer-grouped CV	prediction items disjoint from calibration items
External model / quantizer / size replication	feature families/model class, α=.10, estimators, signed endpoint, anti-leakage and success logic	replication of frozen relationship; no claim of one unchanged coefficient vector across all stages	stage outcomes not used to redesign the prespecified feature family
Held-out concepts/tasks	same protocol plus prospective semantic holdout	evaluates the prespecified relationship on new concept/task domains	held-out concepts/tasks kept outside earlier mechanism-development set
Fresh native-effect recovery	correction predictions fixed before fresh outcomes; no feature redesign	immutable $\hat{D}$ evaluated without predictor refit on the fresh outcomes	96 fresh items; outcome opening after prediction freeze
Decision consequence	zero-threshold rule, correction, and feature design fixed	no threshold changes or post-outcome feature redesign	fresh decision outcomes evaluated after rule freeze
Parsimony	six comparator predictions and primary full-versus-compact comparison fixed	Full predictor does not refit; other comparators fit use only the preceding held-out development stage; all 504 predictions frozen before final parsimony outcomes	fresh recovery/decision outcomes excluded from training; final parsimony outcomes unseen before comparison