Θ: TESTING HEAD EFFECTS THROUGH AVERAGED ACTIVATIONS VIA INJECT-AND-BLOCK

Anonymous authors
Paper under double-blind review

ABSTRACT

Mechanistic interpretability seeks to explain how internal components produce model behavior. After locating a task-relevant attention head, a central question is whether its effect depends on a shared activation direction or on input-dependent structure. This distinction determines how an averaged direction should be interpreted and reused. We introduce **THETA** (Θ), a framework combining mean-direction dependence testing, write geometry, and donor reuse analysis. Its core protocol, **Inject-and-Block**, patches a head, measures the induced residual-stream change, and removes its mean-direction projection, with full-write and random-direction controls. Our contributions are threefold: (i) a controlled protocol connecting an individual head's measured effect to its average write direction; (ii) evidence across five configurations spanning Gemma, GPT-2, Qwen, and Llama that heads exhibit distinct mean-direction removal responses, with write alignment distinguishing the resulting classes; and (iii) a Gemma case study identifying query matching as a condition for reusing writes whose effects persist after mean-direction removal. Same-query donors achieve normalized recovery of 1.00, compared with 0.10 for different-query donors, supported by stronger within-query alignment. THETA connects component-level causal tests with the choice between shared directions and query-matched donor writes.

1 INTRODUCTION

Mechanistic interpretability connects internal components to model behavior. Activation patching locates task-relevant attention heads (Wang et al., 2023); activation steering shifts activations along behavior-related directions (Turner et al., 2023; Rimsky et al., 2024; Zou et al., 2023; Li et al., 2023); and function vectors transfer task information through internal activations (Todd et al., 2024). Connecting these approaches requires a component-level test: how much of an identified head's effect depends on its average direction? This determines when a shared direction captures the head's contribution and when reuse benefits from matching the input.

Consider a head involved in antonym prediction. Patching it changes the model's preference for the correct antonym and induces a residual-stream change, which we call a *write*. Averaging writes gives a candidate summary. Testing it requires connecting this direction to the same head's effect on individual inputs.

Existing studies examine steering reliability across inputs (Tan et al., 2024), predict success from activation geometry (Braun et al., 2025) or layerwise accessibility (Billa, 2026), and learn context-dependent directions (Li et al., 2026). We ask a component-level question with a direct reference: the effect already produced by an identified head. Removing part of its measured write reveals dependence on the candidate average direction.

We introduce **THETA** (*Testing Head Effects Through Averaged Activations*), a framework connecting causal removal tests, write geometry, and donor reuse. Its core protocol, **Inject-and-Block**, patches a head and removes the induced write's projection onto a mean direction estimated from separate examples. Full-write removal and a random-direction control provide paired comparisons (Figure 1). *Static-write heads* lose most of their effect under mean-direction removal; *conditional-write heads* retain substantial effects. Donor experiments then test when a write restores the effect on another example.

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.

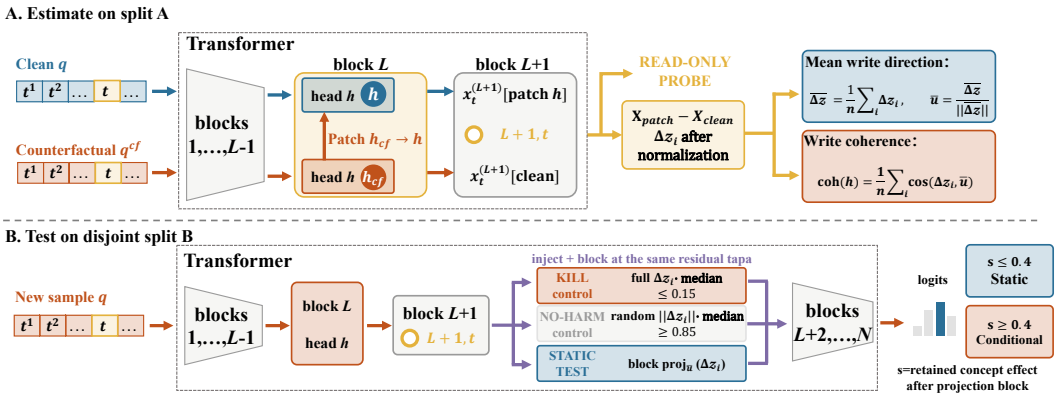

Figure 1: **Inject-and-Block tests mean-direction dependence.** Split A supplies the mean write direction; split B supplies held-out examples for the intervention. At the residual tap, compare full-write removal, random-direction subtraction, and removal of the mean-direction projection. The retained effect $s = s_{\text{mean}}$ assigns static-write heads at $s \leq 0.4$ and conditional-write heads at $s > 0.4$; the boundary value belongs to the static-write class (Equation 4).

Our contributions are threefold: i) **A controlled test of mean-direction dependence.** Inject-and-Block compares projection, full-write, and random-direction removal from the same patched state to test a directional explanation of the head's effect. ii) **Cross-model evidence for distinct removal responses.** Both classes recur across five configurations: antonym prediction in Gemma, Qwen, and Llama, and indirect object identification in GPT-2. *Write coherence*, the alignment of individual writes with their mean direction, distinguishes the classes within each configuration. iii) **Query-dependent write reuse.** In Gemma, matching the donor query restores conditional-write effects, supported by closer alignment within queries than across queries. These contributions give component interpretation a next step after localization: measure mean-direction dependence and test write reuse.

2 Problem Setup

Head effects and per-example writes. We study decoder-only transformers on tasks with a specified contrast, such as antonym prediction versus copying the query word, or the correct versus counterfactual name in indirect object identification (IOI). A query is the task input whose answer is predicted. An example includes that query and its prompt context, such as demonstrations and their ordering. Thus two examples can share a query while differing in context.

Let h denote an attention head at layer L, and t the token position at which its effect is measured. We obtain a patched run by replacing the selected head activation with its counterfactual activation. Write r_i^{clean} and r_i^{patch} for the clean and patched residual states at the downstream measurement tap for example i. The *per-example write* is

$$\Delta z_i = r_i^{\text{patch}} - r_i^{\text{clean}} \in \mathbb{R}^d. \tag{1}$$

The schematics denote this tap by $(L+1, t)$; Appendix E lists the architecture-specific output locations. Measuring the difference at the intervention site includes the effect of normalization on the write. Figure 7 relates the head, token, and residual tap.

Mean write and mean write direction. On an estimation set A of n_A examples, the mean write and its unit direction are

$$\overline{\Delta z} = \frac{1}{n_A}\sum_{i \in A} \Delta z_i, \qquad \bar{u} = \frac{\overline{\Delta z}}{\|\overline{\Delta z}\|}. \tag{2}$$

The mean write retains a magnitude, while $\bar{u}$ specifies the projection axis. Individual examples can have different projection magnitudes along that shared axis.

Retained effect. Let $E_i(r)$ be the downstream effect at residual state r, measured on the task's target logit and normalized so that the clean state has effect zero and the patched state has effect one.

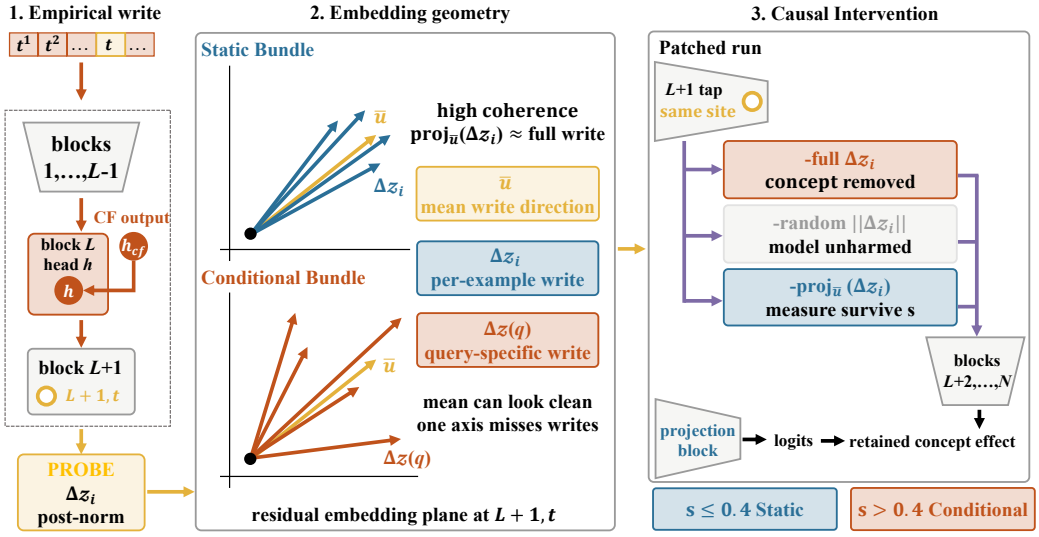

Figure 2: **From write geometry to an intervention on the same head.** Left: measure the per-example write after patching. Middle: a static-write bundle aligns closely with its mean direction, whereas a conditional-write bundle can vary with the query. Right: remove the mean-direction projection and measure the retained effect, with full-write and random-direction controls at the same tap. The geometric patterns motivate the test; the intervention response defines the two classes.

Subtracting a vector at the tap and continuing the forward pass produces the corresponding retained effect. We use the per-head median across evaluation examples, with normalized shares winsorized to $[-1, 2]$. A value near zero denotes removal of the patch effect; a value near one denotes retention. Negative values indicate a reversal relative to the patch, and values above one indicate an effect larger than the reference patch effect.

Only heads with a measurable patch effect enter the analysis. We select them using the magnitude of their patch-to-counterfactual effect, $|\mathrm{PCF}(h)| \geq \epsilon$. The selection floor sets the population on which we compare write patterns. Our primary maps use $\epsilon = 0.02$; the Gemma core analysis uses 0.03.

3 THETA AND THE INJECT-AND-BLOCK PROTOCOL

THETA combines write estimation, Inject-and-Block testing, coherence analysis, and donor comparisons. Inject-and-Block holds the patched example and measurement site fixed while varying the removed vector.

Estimating and testing the direction. We divide examples into an estimation set A and a disjoint evaluation set B. Split A determines $\bar{u}$ through Equation 2. For each example in B, we patch the head and measure its own write Δz_i. The mean-direction component of that write is $\langle \Delta z_i, \bar{u} \rangle \bar{u}$. We compare

$$
\begin{aligned}
s_{\mathrm{mean},i} &= E_i\left(r_i^{\mathrm{patch}} - \langle \Delta z_i, \bar{u} \rangle \bar{u}\right), \\
s_{\mathrm{full},i} &= E_i\left(r_i^{\mathrm{patch}} - \Delta z_i\right), \\
s_{\mathrm{random},i} &= E_i\left(r_i^{\mathrm{patch}} - v_i\right), \qquad \|v_i\| = \|\Delta z_i\|,
\end{aligned}
\tag{3}
$$

where v_i is the random-direction control vector. We denote the per-head medians by s_{mean}, s_{full}, and s_{random}. All three arms use the same patched state and downstream readout. Estimating the direction on A and evaluating it on B tests its use on held-out examples.

Why these controls are paired. Full-write removal checks restoration at the chosen tap; projection removal tests dependence on one axis. Random-direction subtraction tests sensitivity to a perturbation with the full write's norm. Attention-head classification uses median criteria $s_{\mathrm{full}} \leq 0.15$ and

$s_{\text{random}} \geq 0.85$: the complete write restores the reference effect, while an unrelated subtraction largely preserves it. Components outside these control criteria are recorded separately.

The projection coefficient $\langle \Delta z_i, \bar{u} \rangle$ comes from the current example. The test therefore allows the strength of writing along the shared axis to vary across inputs. Donor reintroduction separately measures recovery after adding a chosen write.

Two classes of mean-direction dependence. We summarize the response using

$$\text{class}(h) = \begin{cases} \text{static-write}, & s_{\text{mean}}(h) \leq \tau_s, \\ \text{conditional-write}, & s_{\text{mean}}(h) > \tau_s, \end{cases} \qquad \tau_s = 0.4. \tag{4}$$

We retain continuous s_{mean} values alongside the classes and examine thresholds from 0.2 to 0.6. In figures, "static" and "conditional" abbreviate the two labels; "transition" denotes the intermediate depth band. Figure 2 illustrates the associated write patterns.

Write coherence. The geometric diagnostic is the alignment of per-example writes with their mean direction:

$$\text{coh}(h) = \frac{1}{n_A} \sum_{i \in A} \cos(\Delta z_i, \bar{u}), \qquad \rho = \text{Spearman}\big(\text{coh}(h), s_{\text{mean}}(h)\big). \tag{5}$$

High coherence indicates that the writes share a direction. We evaluate whether this geometric ordering predicts the intervention-defined classes, using the cross-head correlation ρ and the area under the receiver operating characteristic curve (AUC). Here AUC measures how well coherence ranks static-write heads above conditional-write heads: 0.5 corresponds to chance ranking and 1.0 to complete separation. Heads are the units in these comparisons. Once the paired writes have been collected, computing coherence requires no additional downstream blocking runs.

Capturing the empirical write. Architecture determines how an attention output reaches the residual stream. In Gemma-2, post-attention normalization transforms that output. We therefore measure the change after the relevant transformation, using paired clean and patched activations. The normalization comparison in Appendix D demonstrates why this matters: a hand-computed output-space estimate yields retained effect 0.914, compared with 0.000 using the empirical post-normalization write. This makes capture-point selection part of the intervention design.

Testing reuse with donors. Removal identifies dependence on the mean direction. To study reuse, we inject a recipient's own write, a same-query donor's write, an averaged direction, or a different-query donor's write, and compare recovered effects. The same-query arm changes the example while preserving query identity; the different-query arm changes query identity as well. Their comparison tests whether the information needed for recovery is shared across contexts for a given query. Self-reconstruction supplies a reference for the donor comparison. Section 4.2 applies this design to Gemma.

4 Experiments

Models and tasks. We evaluate five model–task configurations. Gemma-2-2b (Gemma Team, 2024) provides the primary antonym-versus-copy task and the detailed donor study. Qwen2.5-7B, Qwen2.5-7B-Instruct (Qwen et al., 2025), and Llama-3.1-8B (Grattafiori et al., 2024) extend the antonym comparison to larger models and additional architectures. GPT-2 small (Radford et al., 2019) uses IOI (Wang et al., 2023), adding a distinct task in which selected heads affect name prediction. Antonym and past-tense tasks follow the function-vector setting of Todd et al. (2024). Past tense supplies a second Gemma task for comparing head recurrence. The base/instruct Qwen pair tests agreement between related checkpoints.

Evaluation design. Cross-model comparisons measure retained effect, its relationship to coherence, and the distribution of classes over depth. Donor recovery and within-query geometry examine the reason for the removal response in Gemma. The MLP extension applies the same intervention design to a different component type (Appendix G). These comparisons answer complementary questions about recurrence across configurations, structure within a task, and reuse across tasks.

Samples and aggregation. The main intervention banks contain 32 items per head, split by even and odd indices into estimation and evaluation sets, with 16 evaluation items per head. Demonstration

Table 1: **Mean-direction dependence across five model–task configurations.** n is the number of selected heads; ρ correlates write coherence with retained effect after mean-direction removal. AUC measures coherence-based separation of the two classes. Class shares use $\tau_s = 0.4$. Parentheses after transition layers give relative depth; the Qwen-Instruct row gives base/instruct class agreement.

Model / task	n	ρ	AUC	Static/cond. (%)	Transition
Gemma-2-2b / antonym	26	-0.81	1.00	$35/65$	L17 (0.65)
GPT-2 small / IOI	13	-0.77	0.95	$38/62$	L9 (0.75)
Qwen2.5-7B / antonym	20	-0.83	1.00	$30/70$	L20 (0.71)
Llama-3.1-8B / antonym	16	-0.86	1.00	$44/56$	L21 (0.67)
Qwen2.5-7B-Instruct / antonym	21	-0.87	0.98	$29/71$	Agreement 0.95

order can vary between examples of the same query. Each head contributes one median retained effect and one coherence score to the cross-head analysis. The expanded Gemma map contains 26 heads; the core set contains 19. We use 5000 permutations for the cross-head replication statistic and 2000 for partial correlations. An initial 12-example pilot is excluded from these results. Appendix E provides the remaining settings.

4.1 Mean-direction dependence separates heads across models

Both static-write and conditional-write heads occur in every evaluated configuration. Table 1 shows static-write shares from 29% to 44%, with the remaining selected heads in the conditional-write class. Coherence is negatively associated with retained effect in all five configurations: ρ ranges from -0.769 to -0.873. The common pattern is that more aligned writes accompany greater effect removal along the mean direction.

The primary Gemma correlation is -0.811 (permutation $p = .0002$), with median retained effects 0.000 under full-write removal and 1.000 under the random-direction control. Mean-direction removal distinguishes heads relative to these restoration and retention outcomes.

Figure 3 places the cross-model and depth comparisons side by side. Panel a compares the correlation estimates and class-separation AUCs across models. Panel b summarizes the Gemma heads by layer band: the static-write fraction decreases from 1.00 in layers up to 13, through 0.80 and 0.40, to 0.00 in layers 20 and above. The bands contain 3, 5, 5, and 13 heads, respectively. This organization connects the response of individual components to their location in the network.

Transitions occur at relative depths 0.65–0.75, with absolute locations from layer 9 in GPT-2 to layer 21 in Llama. Relative depth makes this progression comparable across networks. Thresholds $\tau_s \in [0.2, 0.6]$ give static-write shares of 0.31–0.42 in the analyzed Gemma population; Ashman's $D = 2.93$ summarizes separation of its retained-effect distribution.

The Qwen base and instruct checkpoints share 19 selected heads; 18 retain their class, giving agreement 0.95. The changed assignment lies in the transition band, where continuous retained effects help interpret proximity to the class boundary.

4.2 Query matching restores conditional-write effects

Conditional-write effects in Gemma recover strongly when the donor query matches the recipient query. Figure 4a compares four arms: the recipient's own write (ORACLE), a same-query donor (TWIN), a class-mean direction (MEAN), and a different-query donor (WRONG). Conditional-write recovery is 1.00 for both ORACLE and TWIN, 0.33 for MEAN, and 0.10 for WRONG. The contrast between TWIN and WRONG isolates the practical importance of matching query identity in this donor comparison.

Static-write recovery remains above one with either same-query or different-query donors, including 1.07 for WRONG. Donor matching therefore matters much more for the conditional-write group, connecting the removal-defined classes to a separate reintroduction experiment.

Interpreting recovery. Recovery share expresses the restored concept effect on the experiment's normalized scale; 1 denotes full recovery and values above 1 denote overshoot. Figure 4a and Table 3 summarize arm medians. The normalization was chosen for presentation after inspection of raw

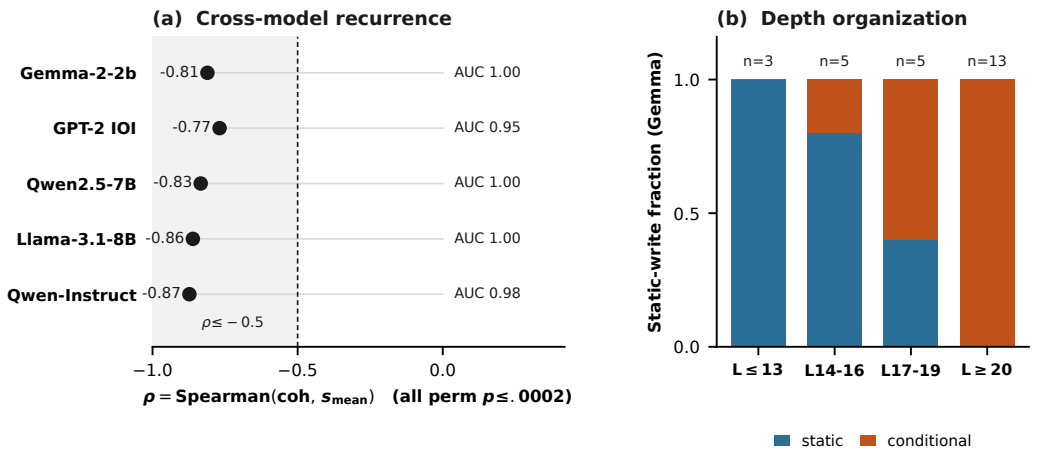

Figure 3: **Cross-model recurrence and depth organization.** (a) Coherence–retained-effect correlations and classification AUCs for the five configurations in Table 1; all permutation $p \leq .0002$. (b) Static-write and conditional-write fractions in four Gemma layer bands. Counts above the bars give the number of heads. Compare the consistently negative association across models with the changing class composition over depth.

magnitudes. Compare donor arms within each group, using ORACLE as the self-reconstruction reference.

Figure 4b presents a separate head-mean steering experiment. Static-write heads reach median recovery 1.25 (IQR $[1.13, 1.54]$), the layer-17 band reaches 0.77, and deep conditional-write heads reach 0.00 (IQR $[0.00, 0.42]$). The two panels address complementary choices: panel a varies donor identity and includes a class-mean direction, while panel b evaluates head-mean steering across depth groups. Their shared pattern is that recovery through an average direction is strongest in the static-write group, whereas matched examples provide effective donors for the conditional-write group.

The geometry behind donor matching. Conditional-write heads have closely aligned writes for the same query and more dispersed writes across queries. Figure 5a shows cosine similarity 0.82 within a query and 0.16 across queries. This directly complements donor recovery: examples sharing a query supply similar write directions and recover the effect, while pooling different queries combines substantially different directions. The mean can therefore obscure structure that is preserved by a matched donor. Figure 2 illustrates this arrangement as query-dependent variation around an overall average.

Reuse across two tasks. On the Gemma past-tense task, the eight heads shared with the antonym set retain their class (8/8; chance agreement 0.62, binomial $P = .022$), and their coherence scores correlate at $r = 0.69$. Recruitment differs between classes: six of nine antonym-selected static-write heads recur on past tense, compared with two of seventeen conditional-write heads (Fisher $p = .008$; Figure 5b). The shared heads preserve their response class, while the different recurrence rates reveal how the two tasks recruit the selected populations. The donor and second-task comparisons thus examine reuse at two levels: changing the example within a query and changing the task in which a head contributes.

4.3 Write coherence identifies different removal responses

Coherence distinguishes static-write and conditional-write heads within each evaluated configuration, with AUC 0.95–1.00 (Table 1). This gives a practical way to rank candidate heads before running the complete blocking comparison. High-ranking heads are candidates for mean-direction dependence; low-ranking heads motivate closer examination of input-specific writes and donor matching.

The scale of coherence differs across configurations. In GPT-2 IOI, static-write coherence spans 0.09–0.66, so a threshold calibrated on Gemma changes the selection. We therefore use coherence to

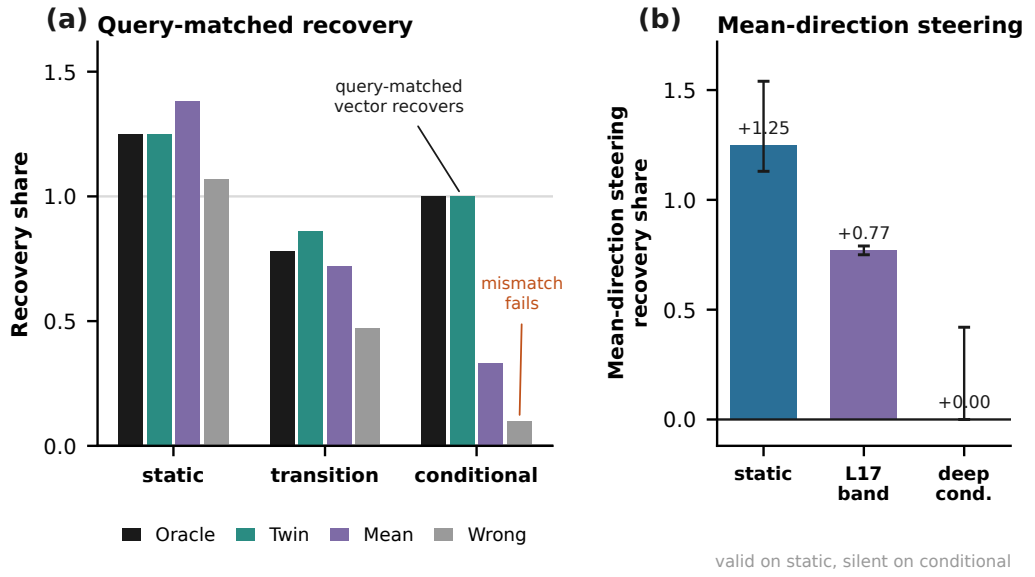

Figure 4: **Query matching and mean-direction steering in Gemma.** (a) Median recovery for own-write, same-query, class-mean, and different-query donors. For conditional-write heads, same-query recovery is 1.00 and different-query recovery is 0.10; static-write recovery remains above one in both arms. (b) A separate head-mean steering experiment, showing medians and IQRs across depth groups. Comparing the panels links the removal response to the choice of average direction or donor write.

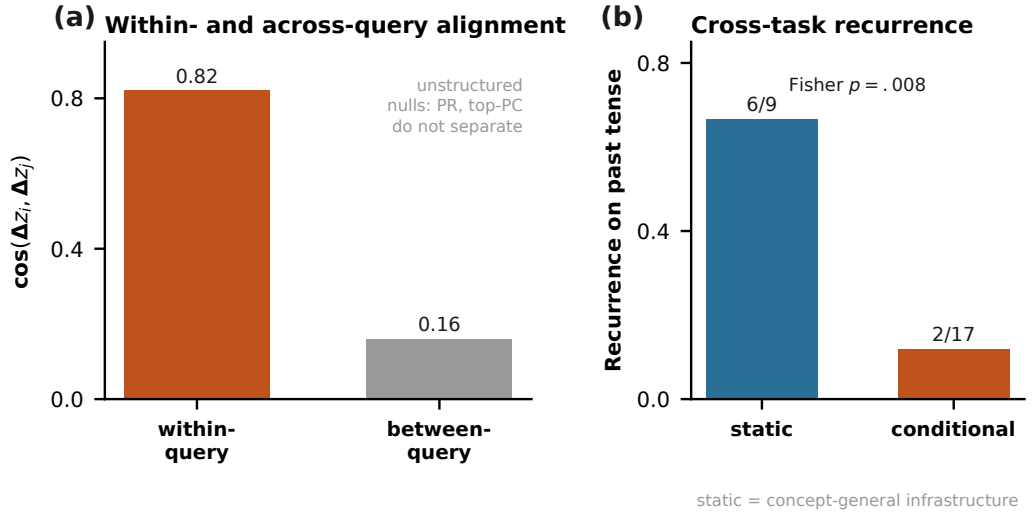

Figure 5: **Query structure and recurrence across two Gemma tasks.** (a) Conditional-write cosine similarity is 0.82 within queries and 0.16 across queries. (b) Recurrence on past tense among antonym-selected heads: $6/9$ static-write heads and $2/17$ conditional-write heads. Compare within- and across-query alignment to interpret donor matching, and compare recurrence rates to assess task recruitment.

order heads within a configuration and calibrate selection against measured removal responses. AUC evaluates this ordering across possible thresholds, explaining how strong class separation can coexist with different absolute score ranges.

Depth organizes the same populations. In Gemma, controlling for layer reduces the coherence–retained-effect correlation from -0.811 to -0.283 ($p = .119$). The geometric and depth summaries therefore capture overlapping structure in these heads. Reporting both makes the diagnostic interpretable: coherence describes the alignment of the writes, while depth locates the components that produce them. The intervention itself supplies the class used in the donor comparisons.

Estimating coherence from a small bank. Four examples per head preserve the sixteen-example ranking at Spearman correlation 0.997 and tercile agreement 0.96. Eight examples give 0.999 and 0.99; twelve give 1.000 and 0.98. A separate split-half comparison gives correlation 0.956 (Appendix F). Small banks thus recover the ordering used to prioritize blocking experiments. Computing the diagnostic requires collecting paired writes and evaluating their alignment.

4.4 Stratifying heads changes predictor associations

Grouping heads by their removal response changes how geometric predictors relate to the contrast between decodability and causal deployment. Decodability measures information recovered by a readout; deployment measures its behavioral effect under intervention. We revisit a screen of more than thirty candidate predictors through five representative diagnostics: attribution patching (AtP), directional agreement, separability, Linear Accessibility Profile (LAP), and read–write cosine. AtP supplies a localization reference (Kramár et al., 2024); directional agreement and separability describe activation geometry (Braun et al., 2025); LAP describes linear accessibility through the output map (Billa, 2026; nostalgebraist, 2020); read–write cosine compares the reading and writing directions.

Figure 6 compares layer-adjusted correlations in pooled, static-enriched, and conditional-enriched populations. Directional agreement changes from -0.070 pooled to $+0.295$ and -0.408 in the two enriched groups. Separability changes from -0.004 to $+0.362$ and -0.236. The near-zero pooled values therefore combine groups with associations of opposite sign. Readers should compare the three bars for each predictor: that comparison reveals the dependence of the association on the sampled head population.

The enriched groups are selected strata used for this diagnostic analysis. A separate held-out check yielded $0/2$ confirmations for two exploratory candidates. The practical result of the stratified comparison is a reporting choice: accompany pooled predictor correlations with the composition of the evaluated heads and their within-group associations. This makes it possible to see when a pooled summary combines different removal responses, and to distinguish a useful descriptive pattern from a predictor that generalizes to new samples.

Using THETA. Select task-relevant heads, collect paired writes, and estimate coherence to prioritize intervention tests. Run Inject-and-Block on separate examples. For heads with substantial retained effect, compare same-query and different-query donors. Report continuous effects alongside class labels to make boundary cases visible. These steps connect component selection, directional summaries, and donor choice.

5 Related Work

Steering reliability. Tan et al. (2024) evaluate how steering vectors generalize across inputs and distribution shifts. Braun et al. (2025) connect agreement among contrastive activation differences to steering effectiveness. Their evaluations relate candidate vectors to the behavioral outcomes of adding them. The component-level question is how much of an identified head's own effect depends on its average write direction. Inject-and-Block addresses this question with a fixed reference effect and matched partial, full, and random-direction removals. The resulting classes can then be compared under donor reintroduction.

Selecting and adapting interventions. LAP predicts steering effectiveness and useful intervention layers from linear accessibility (Billa, 2026). Steering Vector Fields learns an activation-dependent concept score whose gradient supplies a context-dependent direction (Li et al., 2026). These methods guide where to intervene and how to adapt a control vector. Our experiments identify variation in mean-direction dependence among existing attention heads and test when their writes can be reused. The Gemma donor comparison makes query identity an experimentally observable condition for recovery, complementing learned approaches to context-dependent steering.

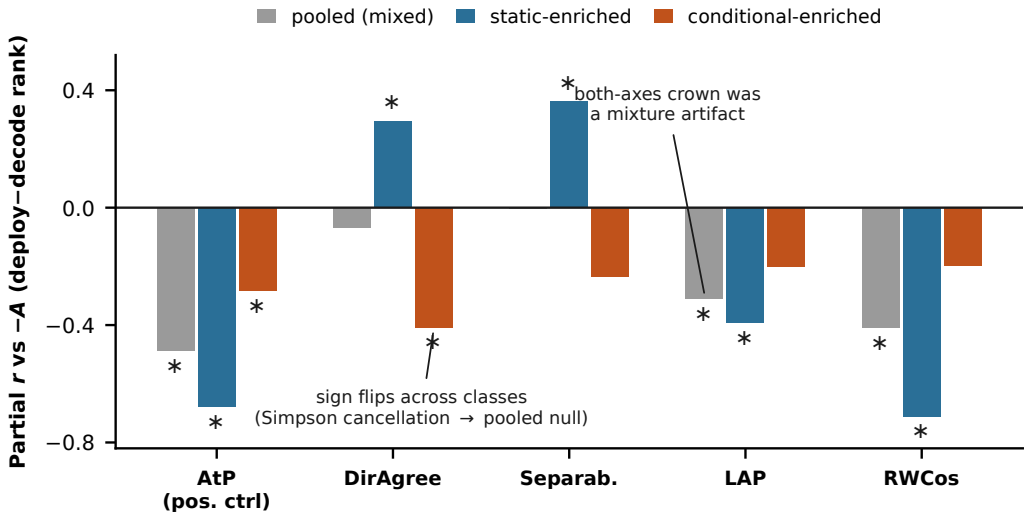

Figure 6: **Predictor associations depend on the head population.** Layer-adjusted correlations with the deployment–decodability contrast for pooled, static-enriched, and conditional-enriched groups; stars mark $p < .05$. Directional agreement and separability have near-zero pooled correlations and opposite signs in the enriched groups. Compare each predictor across groups to see how population composition changes the summary association.

Component localization. Circuit analysis, path patching, and automated discovery locate components affecting outputs (Wang et al., 2023; Goldowsky-Dill et al., 2023; Conmy et al., 2023). Attribution estimates accelerate screening (Kramár et al., 2024; Syed et al., 2023), while causal interventions support model editing (Meng et al., 2022), mediation analysis (Vig et al., 2020), and causal abstraction (Geiger et al., 2021). Inject-and-Block tests the selected component's mean-direction dependence. Its mean-projection arm preserves the example-specific coefficient along the estimated axis.

Representations and component writes. Linear representation studies characterize directions associated with concepts (Park et al., 2024; Marks & Tegmark, 2024; Tigges et al., 2023); sparse autoencoders represent activations through combinations of feature directions (Bricken et al., 2023; Cunningham et al., 2023; Lieberum et al., 2024). Function vectors connect internal activations to transferable task behavior (Todd et al., 2024). Our unit of analysis is the empirical write induced by patching a selected head. Testing the effect of removing its mean-direction component adds a behavioral measurement to the geometric summary. Appendix C compares the experimental questions answered by these lines of work.

6 Conclusion

THETA connects causal testing, write geometry, and donor reuse to explain attention-head effects. Its Inject-and-Block protocol compares partial and full removal of the same empirical write, identifying static-write and conditional-write heads across five model–task configurations. The Gemma experiments give this distinction a concrete consequence: query-matched donors recover conditional-write effects, supported by close within-query alignment. Write coherence helps identify heads with different removal responses. Together, these contributions provide a practical basis for choosing activation summaries and evaluating when component writes can be reused.

Limitations. The cross-model study covers the five configurations in Table 1, while the donor and cross-task analyses focus on Gemma. Head selection and the class threshold determine the evaluated populations; results near the threshold are best interpreted through their continuous retained effects. Extending the donor comparisons to additional models and open-ended generation would test how broadly the observed reuse conditions apply.

Reproducibility Statement

Sections 2–4 describe the intervention, data split, head selection, effect normalization, and statistical comparisons. Appendix E collects the bank sizes, selection thresholds, permutation counts, and architecture-specific capture points. Appendix D illustrates the normalization check, and Appendix F compares the sample sizes used to estimate write coherence. The figure-generation scripts accompany the LaTeX source.

Ethics Statement

This work studies internal activations of existing language models on language tasks used in interpretability research. The experiments involve no human participants or collection of personal data. The interventions are evaluated as tools for analyzing model behavior.

AI Use Statement

AI tools were used to assist manuscript drafting and editing, literature search, and code development, including figure-generation code. The authors take responsibility for the final manuscript and research artifacts.

References

Jayadev Billa. Predicting Where Steering Vectors Succeed. *arXiv preprint arXiv:2604.15557*, 2026.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (Un)Reliability of Steering Vectors in Language Models. In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025. URL https://openreview.net/forum?id=qGCp2AYosf.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards Monosemanticity: Decomposing Language Models with Dictionary Learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards Automated Circuit Discovery for Mechanistic Interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse Autoencoders Find Highly Interpretable Features in Language Models. *arXiv preprint arXiv:2309.08600v1*, 2023.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal Abstractions of Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Gemma Team. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv preprint arXiv:2408.00118*, 2024.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing Model Behavior with Path Patching. *arXiv preprint arXiv:2304.05969*, 2023.

Aaron Grattafiori et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An Efficient and Scalable Method for Localizing LLM Behaviour to Components. *arXiv preprint arXiv:2403.00745*, 2024.

Jiaqian Li, Yanshu Li, and Kuan-Hao Huang. Steering Vector Fields for Context-Aware Inference-Time Control in Large Language Models. *arXiv preprint arXiv:2602.01654*, 2026.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open Sparse Autoencoders Everywhere All at Once on Gemma 2. *arXiv preprint arXiv:2408.05147*, 2024.

Samuel Marks and Max Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2310.06824v3.

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

nostalgebraist. Interpreting GPT: The Logit Lens. *LessWrong*, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Kiho Park, Yo Joong Choe, and Victor Veitch. The Linear Representation Hypothesis and the Geometry of Large Language Models. In *International Conference on Machine Learning (ICML)*, 2024.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115v2*, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models Are Unsupervised Multitask Learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via Contrastive Activation Addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution Patching Outperforms Automated Circuit Discovery. *arXiv preprint arXiv:2310.10348*, 2023.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the Generalisation and Reliability of Steering Vectors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear Representations of Sentiment in Large Language Models. *arXiv preprint arXiv:2310.15154*, 2023.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function Vectors in Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2024.

Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation Addition: Steering Language Models Without Optimization. *arXiv preprint arXiv:2308.10248v1*, 2023.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small. In *International Conference on Learning Representations (ICLR)*, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A THE INTERVENTION UNIT: ONE HEAD, LAYER, AND TOKEN

Figure 7 locates the intervention in a layer-by-token activation lattice. A head is patched at the selected token, the paired runs provide the empirical write, and subtraction is performed at the residual tap used for measurement. The read-only probe records the activation; the intervention changes it before the forward pass continues. Using the same tap connects the measured vector to the downstream effect of removing it. The schematic's $L + 1$ notation identifies this downstream tap, and its survive readout corresponds to s_{mean} in the main text.

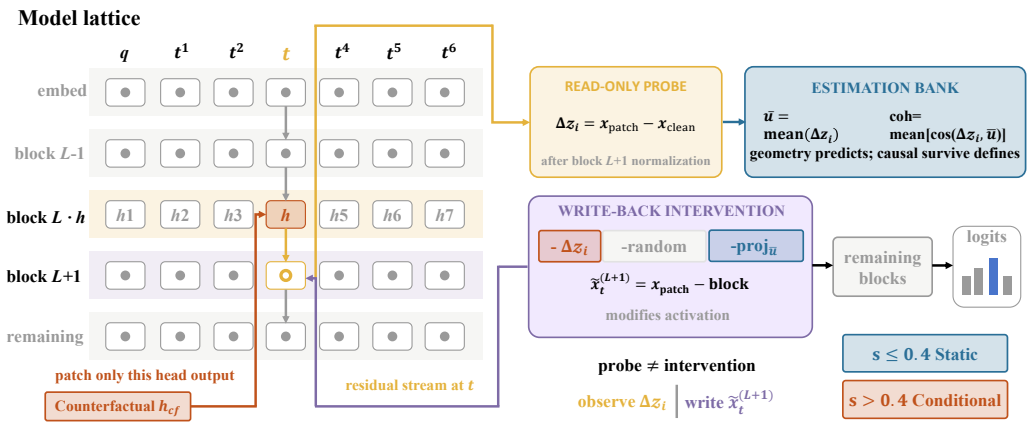

Figure 7: **Locating the head write and the intervention.** The layer-by-token lattice separates the patched head output, the read-only measurement, and the write-back operation at the residual tap. The head's mean-direction dependence is measured by the retained effect after projection removal.

B CROSS-MODEL COMPARISON AND TRANSITION DEPTH

Table 1 and Figure 3 compare the same five model–task configurations. The correlation and AUC summarize the relationship between write geometry and removal response. Transition depth records where the class composition changes along the network. The four listed transition locations are Gemma layer 17 (relative depth 0.65), GPT-2 layer 9 (0.75), Qwen layer 20 (0.71), and Llama layer 21 (0.67). Comparing relative depth places the different absolute layer indices on a common scale. The Qwen-Instruct comparison additionally records agreement with the base checkpoint among shared selected heads.

C Comparison with related methods

Table 2 organizes the closest methods by the question their evaluation answers. THETA starts from an identified component's measured effect, tests mean-direction dependence with Inject-and-Block, and uses donor comparisons to study reuse.

Table 2: **Experimental questions in steering and component interpretation.** The reference effect in Inject-and-Block is produced by the patched head itself. The donor experiment separately measures recovery after reintroduction.

Question	Evaluation
Does steering generalize?	Behavioral effects across inputs and prompt shifts (Tan et al., 2024).
Does agreement predict steering?	Activation-difference geometry compared with mean-vector addition outcomes (Braun et al., 2025).
Where should a layer be steered?	Layerwise output accessibility compared with steering outcomes (Billa, 2026).
How can steering adapt to context?	Learned activation-dependent vector fields evaluated for control (Li et al., 2026).
Which components contribute?	Activation, attribution, and path-patching comparisons (Wang et al., 2023; Kramár et al., 2024).
How much does a head's effect depend on its mean direction?	Mean-projection versus full-write removal from the same patched head, with a random-direction control (Inject-and-Block).
When can a donor write be reused?	Same-query versus different-query recovery across head classes (Gemma donor study).

D Measuring writes after normalization

Post-attention normalization changes the vector that reaches the residual stream. The comparison in Figure 8 evaluates a hand-computed output-space estimate and an empirical difference measured after normalization. The first gives retained effect 0.914, while the empirical write gives 0.000. With the empirical write, the full-write and random-direction controls give 0.000 and 1.000, respectively. The contrast demonstrates a measurement artifact caused by using an estimate that omits the relevant normalization transformation. Capturing paired activations at the intervention site incorporates that transformation directly.

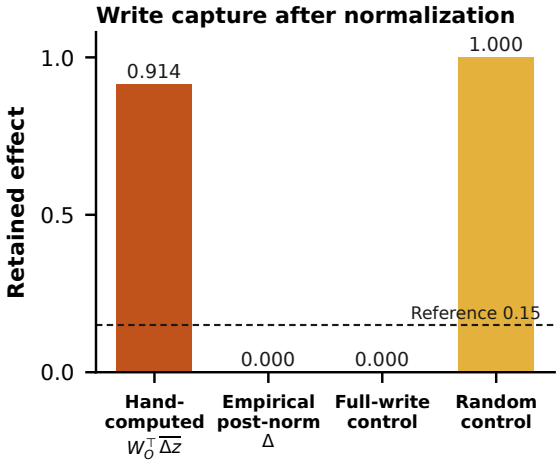

Figure 8: **Normalization changes the measured write.** The hand-computed estimate gives retained effect 0.914; the empirical post-normalization write gives 0.000. The full-write control provides the restoration check, with a reference threshold of 0.15. Compare the estimates to see why the capture point matters.

E EXPERIMENTAL SETTINGS

Selection and readouts. We use class threshold $\tau_s = 0.4$ and examine the range 0.2–0.6. Head-selection floors are 0.03 for the core set and 0.02 for the expanded map. Normalized effect shares are winsorized to $[-1, 2]$ and summarized with per-head medians. Attention-head classification uses full-write median ≤ 0.15 and random-direction median ≥ 0.85. These controls test restoration and sensitivity to an unrelated subtraction, respectively.

Banks and splits. Primary banks use six demonstrations and two orderings. Intervention runs contain 32 items per head, with even indices used for estimation and odd indices for evaluation. The primary and held-out maps use 16 evaluation items per head. The predictor-stratification analysis uses four orderings over 22 words. The smaller configuration uses four examples for coherence and eight evaluation items for blocking. The initial pilot uses 12 examples per head and is excluded from the main analysis.

Statistics. Cross-head correlations use 5000 permutations, and partial correlations use 2000. The decodability analysis uses a ridge-regression probe evaluated by balanced accuracy relative to a permuted baseline. Write coherence uses the estimation bank; retained effects use the evaluation bank. AUC summarizes coherence-based ranking of the intervention-defined head classes.

Capture points. Gemma attention writes are measured as changes in the post-attention-normalization output, and Gemma MLP writes as changes in the post-feedforward-normalization output. GPT-2, Llama, and Qwen use changes in the projection output. These locations follow the transformations through which the component contributes to the residual stream. The empirical measurement supplies the vector for the corresponding subtraction.

F SAMPLE EFFICIENCY

Figure 9 compares coherence rankings from smaller banks with the 16-example reference. At $k = 4$, Spearman rank agreement is 0.997 and tercile agreement is 0.96. At $k = 8$, they are 0.999 and 0.99; at $k = 12$, they are 1.000 and 0.98. Split-half correlation is 0.956 in a separate comparison. A smaller intervention configuration using $k = 4$ for coherence and eight blocking examples reproduces the class map at agreement 1.00 on shared heads. These measurements support using small banks to prioritize components before collecting more extensive intervention results.

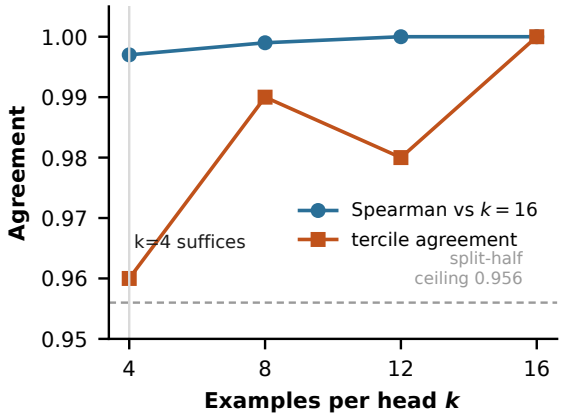

Figure 9: **Coherence ranking with fewer examples.** Compare each estimation-bank size with the 16-example ranking. Four examples give Spearman agreement 0.997 and tercile agreement 0.96. The separately measured split-half correlation is 0.956.

G Extension to MLP components

Across eight Gemma MLP components, write coherence is negatively associated with retained effect ($\rho = -0.833$, permutation $p = .012$). Figure 10 shows the individual components: high-coherence MLPs retain less effect after mean-direction removal, while low-coherence MLPs retain more. Static-write components occur at layers 13–14, before the attention transition at layer 17, and the subsequent components have larger retained effects.

The random-direction control ranges from 0.51 to 1.45 across these MLPs. MLP15 is marked in grey because it is also sensitive to random-direction subtraction. Its projection-removal response should therefore be read together with that control response. The plot retains all eight components to show the relationship between write geometry, layer position, and sensitivity to the intervention.

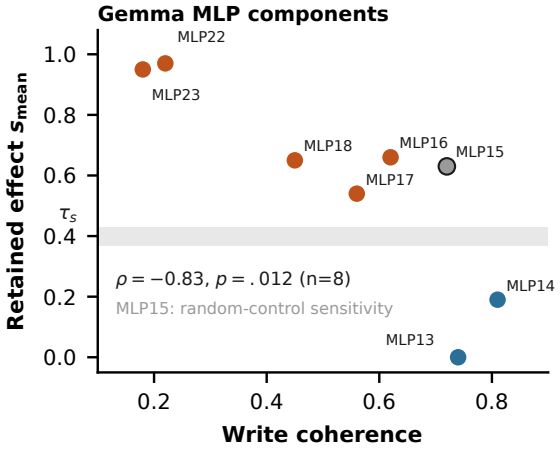

Figure 10: **Write coherence and retained effect in Gemma MLPs.** Each point is one of eight components ($\rho = -0.83$, permutation $p = .012$). The shaded band marks the classification threshold. Grey identifies MLP15, which also shows sensitivity to random-direction subtraction.

H Donor recovery and checkpoint comparison

Table 3 gives the donor-arm medians behind Figure 4a. ORACLE uses the recipient's own write; TWIN uses a same-query donor; MEAN uses the class-mean direction; WRONG uses a different-query donor. The transition group summarizes the intermediate depth band. Compare TWIN and WRONG within each group to see how query matching changes recovery.

Table 3: **Donor recovery in Gemma.** Median normalized recovery shares for the four reintroduction arms. The count column reproduces the donor-analysis counts. A value of one denotes full recovery on the experiment's scale.

Group	Count	ORACLE	TWIN	MEAN	WRONG
Static-write	104	+1.25	+1.25	+1.38	+1.07
Transition	24	+0.78	+0.86	+0.72	+0.47
Conditional-write	154	+1.00	+1.00	+0.33	+0.10

For Qwen2.5-7B and Qwen2.5-7B-Instruct, 18 of 19 shared selected heads retain their class. The changed assignment occurs in the transition band, where retained effects lie closer to the boundary. This comparison measures agreement among shared heads in the two checkpoints and complements the cross-model statistics in Table 1.