LINEAR READABILITY DOES NOT GUARANTEE WRITABILITY: EVIDENCE FROM CONTROLLED COUNTERFACTUAL RECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Linear probing reveals information encoded in large language models, but whether probe-derived directions can reliably steer model decisions remains unclear. To investigate this question, we carefully design a controlled experiment on tasks constructed from ALFWorld and WebShop trajectories, whose action lists allow us to define explicit state changes and determine the corresponding answers. In the experiments, we combine the source input's queried state with the target input's decision rule and answer mapping, which we term counterfactual recomposition, to test whether steering supports the intended recomputation rather than merely copying an answer or source plan. Simultaneously, we learn and validate a reference steering subspace to ensure that steering can produce the required counterfactual answer at the selected location. This reference enables a controlled comparison with probe-derived steering in the complement space under fixed intervention conditions. Across four models and both environments, probing accuracy remains near-perfect in the complement space, while steering with direct probe directions, covariance-transformed directions, or combinations of multiple probe directions achieves near-zero counterfactual recomposition accuracy. These results reveal a critical finding: high readability alone does not guarantee writability through probe-derived interventions even with effective steering available, which provides valuable insight for developing reliable steering methods.

1 INTRODUCTION

Linear probing and steering represent two fundamental activation-based paradigms that operate on the internal representations of large language models from read and write perspectives, respectively. Probing extracts model activations to decode task-relevant signals, while steering modifies activations to induce corresponding behavioral shifts. These methods have been used to enhance model performance, for example, to filter unreliable answers in selective question answering (Bao et al., 2025), reduce hallucinations in vision-language generation (Su et al., 2025), and improve formal reasoning across languages (Maraia et al., 2026).

However, the relationship between probing and steering is insufficiently explored. Prior studies have observed that a direction achieving high probing accuracy does not necessarily yield an effective steering axis, suggesting that readability does not guarantee writability. Information decoded by a probe need not be causally used in the computation of the model (Elazar et al., 2021; Sharma et al., 2026), and interventions optimized directly for behavioral effects can outperform those derived from probe classifiers (Arora et al., 2024; Khorasani et al., 2026). More directly, Lepori et al. (2024) find that optimizing activations to change probe predictions can fail to produce the expected model outputs. These findings highlight the discrepancy between reading and writing, while it remains unclear to what extent directions obtained by linear probing support steering interventions that reliably make the model answer according to a changed task state.

To address this question, we carefully design our experiments to test whether probe-derived directions make the model answer according to a changed task state when effective steering is achievable at the same location. For a controlled comparison, these interventions are conducted with fixed layer, position, sample pairs, and replacement operation within each model and environment. We

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.

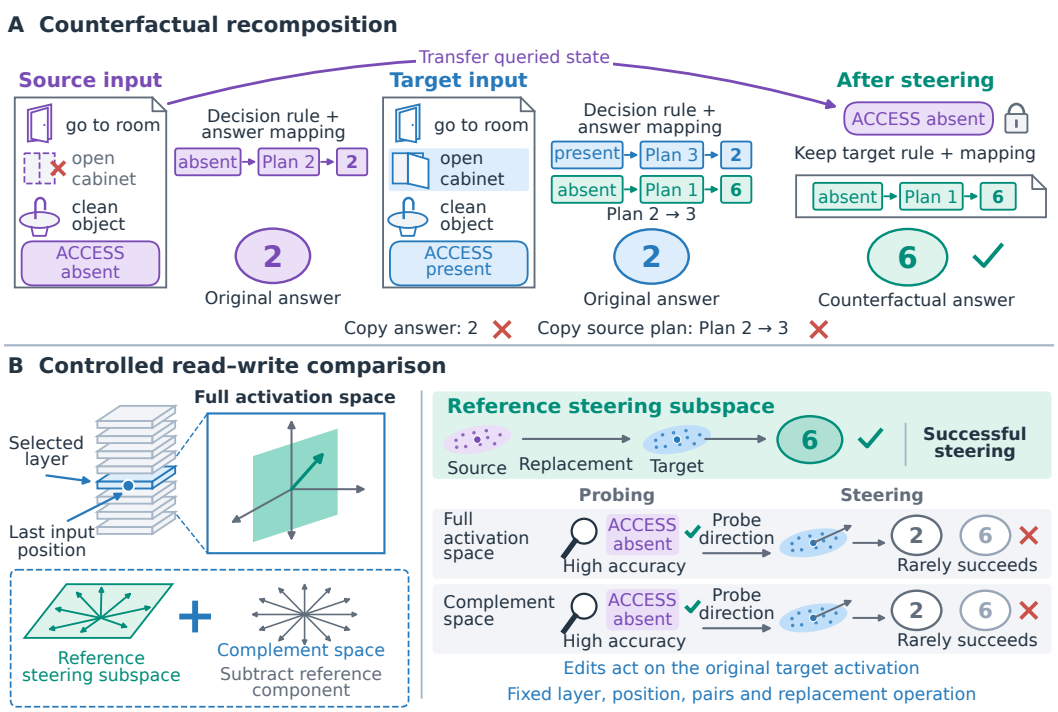


Figure 1: Controlled recomposition and the read–write comparison. **(A)** Removing the opening action changes ACCESS, but both original answers are 2. Combining the source state with the target decision rule and answer mapping requires Plan 1 and answer 6; copying an original answer or the source plan instead yields 2 or 3, respectively. **(B)** State remains readable outside a validated steering subspace, yet direct probe directions rarely support recomputation. All edits act on the original target representation at a fixed layer and position.

use ALFWorld (Shridhar et al., 2021) and WebShop (Yao et al., 2022) trajectories to construct our controlled tasks, as they provide concrete task states and explicit action lists. In each task, we define a queried state indicating whether a particular type of action appears in the list; a decision rule maps this state to a plan label and an answer mapping converts the label to the final answer. Effective steering should make the model recompute the answer from the changed queried state using the target decision rule and answer mapping. This controlled setting allows us to evaluate steering against a clearly specified answer rather than an unspecified change in model output.

Inspired by the intervention-based evaluation (Geiger et al., 2021), we design a counterfactual recomposition experiment, which tests whether steering makes the model combine the queried state of the source input with the decision rule and answer mapping of the target to recompute the answer. To exclude copying, we construct source and target inputs that originally share the same answer, while counterfactual recomposition requires a different answer that cannot be obtained by reusing the source plan. Thus, producing this counterfactual answer provides evidence for the intended recomputation beyond simple answer or source-plan copying. Figure 1(A) illustrates this construction.

Building on the subspace learning methods (Geiger et al., 2024; Wu et al., 2023), we learn and validate a low-dimensional reference steering subspace within the full activation space, demonstrating that steering can produce the required counterfactual answer at the selected location. To test whether readable information outside this subspace supports steering, we fit probes after subtracting the reference steering subspace component from activations and use the probe-derived directions to perform steering interventions. We find that linear probing still predicts the queried state with near-perfect accuracy in this complement space, but steering along the resulting probe directions rarely produces the required counterfactual answer (see Figure 1(B)). To test whether this separation depends on using individual probe coefficients directly as steering directions, we also examine two further probe-derived constructions: covariance-transformed directions and combinations of multiple probe directions. Despite improving counterfactual recomposition accuracy in the full activation space,

these constructions remain ineffective in the complement space. Consequently, our results show that high readability does not guarantee that the tested probe-derived directions provide writability, even when effective steering is achievable at the same location.

Our main contributions are as follows:

- **Counterfactual Recomposition Beyond Copying:** We construct controlled counterfactual recomposition tasks that distinguish state-dependent recomputation from answer and source-plan copying, providing an explicit criterion for evaluating steering effectiveness.
- **Controlled Read-Write Comparison:** We establish effective steering through a reference subspace and compare probe-derived interventions under the same intervention conditions, separating whether recomputation is achievable from whether probe-derived directions can support it.
- **A Persistent Read-Write Gap:** Across different models and environments, we show that near-perfect probing accuracy persists outside the reference subspace, while the tested probe-derived interventions achieve near-zero counterfactual recomposition accuracy, demonstrating that high readability does not guarantee writability across models even when effective steering is achievable.

2 Related Work

Probing and the use of readable information. Probing measures readability by predicting task variables from model activations. Control tasks introduced by Hewitt & Liang (2019) distinguish information learned by the probe from information supplied by the model, while Ravichander et al. (2021) show that high probing accuracy can persist for properties that are irrelevant to the model's task. Information-removal studies test whether models use readable information by removing it from activations and measuring changes in model predictions (Elazar et al., 2021). Such tests require checking both the removal of target information and the preservation of other information, motivating more precise removal methods and reliability evaluations (Dobrzeniecka et al., 2025; Canby et al., 2025). In controlled sequence tasks, Sharma et al. (2026) further show that removing probe directions for variables with high readability can have little behavioral effect. Using sparse autoencoder features, Tiwari et al. (2026) find that removing features selected by probe alignment alone changes probe scores more, but model outputs less, than removing features selected jointly by probe alignment and output sensitivity. In agentic tasks, probing environment states and plans also enables comparisons between internal representations and observed actions (Arghal et al., 2026). These studies distinguish readability from the model's use of information; whether that information supports steering requires a further behavioral test.

From probing to steering. Steering directions can be constructed from average activation differences between paired examples (Rimsky et al., 2024) or from class means at attention heads selected by probing (Li et al., 2023). In CausalGym, Arora et al. (2024) directly compare probe coefficients with behaviorally optimized directions under a common replacement operation. In controlled arithmetic tasks, Lepori et al. (2024) evaluate activation edits optimized against fixed probes: the edits change probe predictions but rarely produce the expected downstream answers. Strategic games also exhibit limited or unreliable probe-derived steering (Sobotka et al., 2026). The choice of direction matters: Park et al. (2024) connect probing and steering under explicit assumptions about linear representations and the inner product, while FishBack constructs directions from output geometry (Wang et al., 2026) and inference-time causal probing optimizes target-versus-alternative output scores (Khorasani et al., 2026). Our comparison tests whether readability outside a validated reference steering subspace supports probe-derived steering, including covariance transformations and combinations of probe directions, with the layer, position, and replacement operation held fixed.

Counterfactual evaluation of writability. Causal abstraction evaluates neural interventions against the output of a specified computation after changing a target variable (Geiger et al., 2021). Building on this criterion, Geiger et al. (2024) introduce Distributed Alignment Search (DAS), which learns subspaces by replacing part of a target activation with the corresponding part of a source activation and optimizing for the prescribed counterfactual output. Boundless DAS (Wu et al., 2023) also learns the subspace size. These methods motivate our reference steering subspace, which establishes

writability independently of probing accuracy. Evaluating the intended change also requires checking what should remain unchanged: Huang et al. (2024) evaluate both attribute transfer and preservation of other attributes in RAVEL. In our setting, the transferred state must be processed using the target decision rule and answer mapping to produce an answer different from both original answers, excluding answer copying. Sources with the same queried state must instead preserve the target answer, providing a complementary check against unconditional branch switching.

3 Controlled Read-Write Analysis

Our analysis examines whether directions obtained by linear probing can also make the model answer according to a changed task state. We use counterfactual recomposition to evaluate probe-derived steering against a validated reference steering subspace under the same intervention conditions. We evaluate probe-derived constructions using direct probe directions, covariance-transformed directions, and combinations of multiple probe directions in both the full activation space and complement space.

3.1 Controlled task

Each input contains an action list, a decision rule, and an answer mapping, which are formalized as $x = (S, G, C)$. The action list denotes the binary state $S \in \{0,1\}^K$ for K action roles, with $S_k = 1$ if and only if an action of role k is present. The decision rule $G : \{0,1\}^K \to \mathcal{Q}$ maps states to plan labels in $\mathcal{Q} = \{q_1, \ldots, q_M\}$ using only one designated coordinate $j(G) \in \{1, \ldots, K\}$. The value $S_{j(G)}$ is called the queried state; the rule selects one label when it is 1 and a different label when it is 0. An answer mapping $C : \mathcal{Q} \to \mathcal{A}$ assigns each label a distinct answer from the answer set $\mathcal{A}$. The selected plan label Q and answer A are

$$Q = G(S), \qquad A = C(Q). \tag{1}$$

The prompt expresses S through the action list and supplies G and C as a two-branch decision rule and an M-row answer mapping. This separation lets us change the queried state while retaining the rule and answer mapping that determine how the target should respond.

We instantiate the tasks with action lists derived from successful ALFWorld (Shridhar et al., 2021) and WebShop (Yao et al., 2022) trajectories, using $K = M = 4$ and $\mathcal{A} = \{0, \ldots, 9\}$. The state coordinates are VIEW, ACCESS, MODIFY, and SUBMIT; each present role contributes exactly one action. In ALFWorld these roles correspond to navigation or inspection, object access, modification, and placement. In WebShop they correspond to search or navigation, product access, option selection, and purchase. The shared roles permit the same controlled construction with environment-specific actions; the decision rules and answer mappings are supplied by our task generator.

3.2 Counterfactual recomposition

For a target $x = (S, G, C)$ and a source $x' = (S', G', C')$ whose decision rules check the same coordinate $j = j(G) = j(G')$, counterfactual recomposition requires the source's queried state to be processed using G and C. Let S^{cf} be the target state S with only coordinate j replaced by the source value S'_j. The counterfactual prompt $x^{\mathrm{cf}} = (S^{\mathrm{cf}}, G, C)$ changes the target action list accordingly and defines the required plan and answer:

$$Q^{\mathrm{cf}} = G(S^{\mathrm{cf}}), \qquad A^{\mathrm{cf}} = C(Q^{\mathrm{cf}}). \tag{2}$$

During steering, the target prompt itself is unchanged; the intervention must make the model produce A^{cf} by modifying its activation.

We construct opposite-value sources so that copying their answer or plan cannot satisfy this requirement. Starting from the target, we flip only the queried state coordinate and replace the plan on the newly selected branch of the source rule with a label distinct from both target rule labels. Write $Q' = G'(S')$ for the source plan. The resulting plans satisfy

$$Q \neq Q', \qquad Q \neq Q^{\mathrm{cf}}, \qquad Q' \neq Q^{\mathrm{cf}}. \tag{3}$$

The source answer mapping is obtained from C by swapping the answers assigned to Q and Q'. Thus, neither original answer equals the required counterfactual answer. Reusing Q' with the target answer

mapping also fails because $C(Q') \neq C(Q^{\mathrm{cf}})$.

$$C'(Q') = C(Q) \neq C(Q^{\mathrm{cf}}). \tag{4}$$

Figure 1(A) illustrates these distinct outcomes. An additional same-value source control checks for unconditional branch switching (Appendix E).

3.3 Steering operation and reference steering subspace

Let $h_p^\ell(x) \in \mathbb{R}^d$ be the d-dimensional residual-stream activation at layer ℓ and prompt position p. All interventions use the final prompt position p. Let x^v denote input x with canonical ($v = \mathrm{C}$) or reversed ($v = \mathrm{R}$) rule-row order, with v_s, v_t denoting the source and target orders. Write $h' = h_p^\ell((x')^{v_s})$ and $h = h_p^\ell(x^{v_t})$. Given an orthonormal basis $U \in \mathbb{R}^{d\times r}$ for an r-dimensional candidate steering subspace, we apply

$$\widetilde{h} = h + UU^\top(h' - h)\,. \tag{5}$$

This replaces the target activation's component in the candidate subspace with the source component and leaves the orthogonal component unchanged. The replacement coefficient is fixed at one for every construction. For model F, counterfactual recomposition accuracy is the empirical success rate over the eligible pairs $\mathcal{E}_F$ defined in Section 4.1:

$$\mathrm{Acc}_{\mathrm{CF}}^{v_s\to v_t}(U) = \Pr\Big[F\Big(x^{v_t}; \mathrm{do}(h = \widetilde{h})\Big) = A^{\mathrm{cf}} \;\Big|\; (x', x) \in \mathcal{E}_F\Big]\,. \tag{6}$$

Here F selects the highest-scoring answer in the target answer mapping, and do denotes the activation replacement in Eq. 5.

To ensure that steering can produce the required counterfactual answer, we learn a reference steering subspace independently of probing accuracy. We select the layer on a prespecified subset of selection data using minimal pairs that share the decision rule and answer mapping and differ only in the queried state. For these localization pairs (x', x), let $\mathbf{z}_F$ denote the model's logits over the M target answer candidates. For each intervention layer ℓ and later layer $k > \ell$, we compute

$$\begin{aligned}\mathbf{z}_\ell &= \mathbf{z}_F\big(x; h_p^\ell \leftarrow h_p^\ell(x')\big)\,,\\ \mathbf{z}_{\ell,k} &= \mathbf{z}_F\big(x; h_p^\ell \leftarrow h_p^\ell(x'),\ h_p^k \leftarrow h_p^k(x)\big)\,,\end{aligned} \tag{7}$$

where arrows denote activation replacements during the target forward pass, and $h_p^k(x)$ is taken from the unmodified target. The first intervention tests whether full source replacement changes the decision; the second tests whether restoring the target activation later removes that change. Layer pairs are screened and ranked by prespecified behavioral margins, including these effects and answer-mapping controls, and the selected ℓ is used for all subsequent steering. The screening criteria and ranking are specified in Appendix B.

For each model and environment, the base model is frozen and a single reference basis is learned at the selected layer from training data. To learn its rank, we use an ordered basis $B \in \mathbb{R}^{d\times m}$ of m orthonormal directions, with column weights a_i controlled by a learned boundary $\rho \in [0, m]$:

$$a_i = \sigma\left(\frac{\rho - (i - \frac{1}{2})}{\tau}\right), \qquad P = B\,\mathrm{diag}(a_1, \ldots, a_m)B^\top, \qquad r_{\mathrm{soft}} = \sum_{i=1}^{m} a_i, \tag{8}$$

where σ is the sigmoid function and τ controls the smoothness of the boundary. Here r_{soft} estimates the rank continuously, and P is the weighted replacement matrix used in place of $UU^\top$ during fitting. Applying this replacement from x' to x produces answer logits $\mathbf{z}_{s\to t}^P$; applying it from x to x^{cf} produces $\mathbf{z}_{t\to \mathrm{cf}}^P$. We minimize

$$\begin{aligned}\mathcal{L}(B, \rho) = \mathbb{E}_{\mathrm{train}}\big[&\mathrm{CE}(\mathbf{z}_{s\to t}^P, A^{\mathrm{cf}}) + \mathrm{CE}(\mathbf{z}_{t\to \mathrm{cf}}^P, A)\big]\\ &+ \lambda_E \mathcal{E}(P) + \lambda_r r_{\mathrm{soft}},\end{aligned} \tag{9}$$

$$\mathcal{E}(P) = \mathbb{E}_{\mathrm{train}}\left[\frac{\|P(h' - h)\|_2^2 + \|P(h - h^{\mathrm{cf}})\|_2^2}{2\|h\|_2^2 + \epsilon}\right], \tag{10}$$

where the expectation covers paired training examples under both rule presentations, $h^{\mathrm{cf}} = h_p^\ell((x^{\mathrm{cf}})^{v_t})$ is the direct counterfactual activation at the same location and target rule-row order,

and ϵ is a numerical stabilizer. The cross-entropy loss CE uses the M answers in the target mapping. The forward term trains counterfactual recomposition, while the reverse term trains recovery of the original target answer from the direct counterfactual. The weights λ_E and λ_r control the replacement-energy and effective-rank penalties, respectively.

After optimization, we retain columns of B with $a_i \geq \frac{1}{2}$ to form U and refine this orthonormal basis with $P = UU^\top$, keeping the energy penalty and dropping the rank penalty. Rank and restart are selected using the less favorable rule presentation on the prespecified selection subset. Hereafter, U denotes this frozen reference basis, shared across queried coordinates, decision rules, plan labels, answer mappings, and presentations; its held-out counterfactual accuracy provides the behavioral reference for the subsequent comparison with probe-derived steering. Appendix B gives the initialization, optimization schedule, and candidate-ranking details needed to reproduce this fit. Additional checks of the learned reference are reported in Appendix D.1.

3.4 Linear probing and direct probe-derived steering

With an effective reference at the selected location, we test whether directions obtained by probing also support counterfactual recomposition. We fit linear ridge classifiers to predict the queried state $S_{j(G)}$. One probe uses the full activation; another uses the complement-space activation obtained by subtracting the reference steering subspace component:

$$h_\perp = (I - UU^\top)h_p^\ell(x). \tag{11}$$

Here I is the identity matrix. The complement-space probe tests whether state information remains readable outside the validated reference steering subspace.

Features are standardized using training-set statistics, and probe regularization is chosen on selection data. For the primary steering direction, we map the standardized-feature coefficients w_{std} back to the original activation coordinates and normalize them:

$$D = \text{diag}(s_1, \ldots, s_d), \qquad w = D^{-1}w_{\text{std}}, \qquad u = \frac{w}{\|w\|_2}, \qquad U_{\text{probe}} = [u], \tag{12}$$

where s_i is the training standard deviation of feature i and U_{probe} is the one-column basis containing u. For the complement-space probe, w is replaced by $(I - UU^\top)w$ before normalization. Probes are fitted on canonical training prompts. Readability is reported as held-out balanced accuracy of a separate ridge classifier fitted to activations projected onto each frozen probe direction, using training and selection data. Using U_{probe} in Eq. 5 gives a rank-one steering intervention, following the coefficient-based comparison in CausalGym (Arora et al., 2024). Both constructions apply Eq. 5 to the original target activation, not to the complement-space activation $h_\perp$.

3.5 Covariance transformations and combinations of probe directions

The primary comparison makes two choices: it uses normalized probe coefficients as the steering direction and restricts the intervention to one dimension. We examine these choices through two further probe-derived constructions: covariance transformations and combinations of multiple probe directions, respectively.

Motivated by the role of geometry in relating probing to steering (Park et al., 2024), we test a covariance transformation using the training activations. For a probe coefficient w in the original activation coordinates, the transformed direction is

$$u_{\text{cov}} = \frac{\widehat{\Sigma}w}{\|\widehat{\Sigma}w\|_2}, \tag{13}$$

where $\widehat{\Sigma}$ is the covariance matrix estimated from the centered training activations used for that probe. This construction changes the direction while retaining a rank-one intervention.

To test whether steering benefits from multiple state-predictive directions, we combine coefficients fitted to different state variables and query conditions. The candidate pool contains probes for each state coordinate under each rule presentation, queried-state probes fitted separately for each queried coordinate, and a pooled queried-state probe fitted across coordinates. The canonical state-coordinate probes use both base tasks and controlled changes to states, action wording, decision rules,

and answer mappings. Reversed state-coordinate probes use base tasks with reordered rule rows; coordinate-conditioned and pooled queried-state probes use base tasks under both presentations.

Given the resulting J coefficient vectors $w_1, \ldots, w_J$ in the original activation coordinates, we normalize each column and compute a singular value decomposition:

$$W = \left[\frac{w_1}{\|w_1\|_2}, \ldots, \frac{w_J}{\|w_J\|_2} \right] = L \Lambda R^\top. \tag{14}$$

Here L and R contain the left and right singular vectors, and Λ is diagonal with singular values in descending order. Writing $r_{\text{ref}} = \text{rank}(U)$ for the reference rank, we retain

$$r_{\text{multi}} = \min(r_{\text{ref}}, \text{rank}_{\text{num}}(W)), \qquad U_{\text{multi}} = L_{:,1:r_{\text{multi}}}, \tag{15}$$

where rank_{num} counts singular values above the numerical threshold. The first r_{multi} columns of L form the orthonormal basis U_{multi}, with dimension bounded by the reference rank.

We also combine covariance-transformed directions. For each coefficient w_i, let $\widehat{\Sigma}_i$ be the covariance matrix of its training subset, giving

$$W_{\text{cov}} = \left[\frac{\widehat{\Sigma}_1 w_1}{\|\widehat{\Sigma}_1 w_1\|_2}, \ldots, \frac{\widehat{\Sigma}_J w_J}{\|\widehat{\Sigma}_J w_J\|_2} \right]. \tag{16}$$

Applying the same decomposition and rank truncation to W_{cov} produces the combined covariance-transformed steering basis.

We repeat all constructions using complement-space training activations and project the resulting multi-direction bases back into the complement space before orthonormalizing them. All resulting directions and bases act on the original target activation through Eq. 5. Appendix F gives the numerical rank threshold and implementation details.

4 Experiments

4.1 Experimental setup

We evaluate the controlled tasks in Section 3 on four instruction-tuned models: Qwen2.5-14B-Instruct (Qwen Team, 2025), Qwen3-14B (Yang et al., 2025), Phi-4 (Abdin et al., 2024), and Ministral-3-14B-Instruct-2512 (Liu et al., 2026). All of them are tested on ALFWorld- and WebShop-derived tasks with non-thinking, text-only instruction behavior where applicable, yielding eight fixed model–environment conditions. For each environment, the shared task generator provides 160 training groups for learning the reference steering subspace, fitting probes, and estimating activation covariances; 32 selection groups for choosing intervention and probe settings; and 416 final groups for evaluating the selected constructions. Underlying tasks and answer mappings are disjoint across splits. Counterfactual recomposition is evaluated on tasks the unmodified model can solve. An opposite-value pair is eligible for model F when its source, target, and direct counterfactual prompts all receive their correct answers under both semantically equivalent rule-row orders (Appendix D.2), scoring each prompt over the M answers in its own mapping. The eligible pairs form $\mathcal{E}_F$ and are fixed before steering evaluation. Within each environment, every model is evaluated on the same 416 candidate final groups. Eligible n in Table 1 counts the groups that pass this rule and varies across models according to their unmodified answers; all steering constructions for a given model use the same eligible groups under both rule-row orders.

For each model–environment condition, we learn a reference steering subspace from training data. A fixed subset of 24 of the 32 selection groups is used to select the intervention layer, reference rank, and optimization restart. The selected reference steering subspaces have rank 8 or 9 (Table 4). All steering constructions use the selected layer, the final prompt position, and the replacement operation in Eq. 5 with coefficient one. Optimization settings are given in Appendix B.

At the selected location, we fit linear probes on training activations in the full activation space and complement space. Probe regularization is selected using the full 32-group selection split. The normalized probe coefficients define the direct probe-derived steering directions. In subsequent post-hoc analyses, we evaluate covariance-transformed directions and combinations of multiple probe

directions in both spaces. These constructions use training data for probe fitting and covariance estimation and selection data for regularization, with the combined rank capped by the reference rank, as defined in Section 3.5. The final groups evaluate the selected reference steering subspace and all probe-derived constructions. Probing uses all final groups without the steering eligibility filter and reports the lower balanced accuracy across its canonical and reversed evaluation sets (Appendix C). We report counterfactual recomposition accuracy averaged over the two same-presentation steering conditions: canonical source with canonical target, and reversed source with reversed target. Cross-presentation experiments are reported in Appendix D.2.

4.2 Reference steering and direct probe-derived steering

Table 1 compares readability and writability at the same intervention location. The reference steering subspace achieves high counterfactual recomposition accuracy across all eight model–environment conditions, using one shared subspace per condition across queried states, decision rules, plan labels, answer mappings, and presentations. This accuracy measures agreement with the answer obtained by combining the source queried state with the target decision rule and answer mapping. Since the required answer differs from both an original answer and the answer obtained by reusing the source plan, successful interventions provide evidence for counterfactual recomposition beyond these copying strategies. The reference thus establishes writability against a task-specific behavioral criterion and provides a positive control for examining whether directions obtained by probing achieve the same change.

Table 1: Probing and counterfactual recomposition accuracies under shared intervention conditions. Eligible n counts steering pairs drawn from 416 final groups per environment. Probing columns report the lower balanced accuracy across canonical and reversed rule-row orders. Reference and probe-derived steering columns report mean counterfactual recomposition accuracy across canonical-to-canonical and reversed-to-reversed source–target conditions. Full and Comp. denote the full activation space and complement space.

				Probing (%)		Probe-derived steering (%)	
Environment	Model	Eligible n	Reference (%)	Full	Comp.	Full	Comp.
ALFWorld	Qwen2.5	406	94.8	100.0	100.0	0.0	0.0
	Phi-4	415	98.2	100.0	100.0	0.0	0.0
	Ministral 3	378	88.5	100.0	100.0	0.0	0.0
	Qwen3	311	94.9	99.7	100.0	0.8	0.3
WebShop	Qwen2.5	285	90.4	100.0	100.0	0.2	0.0
	Phi-4	402	98.9	100.0	100.0	0.0	0.0
	Ministral 3	304	96.2	100.0	100.0	0.5	0.5
	Qwen3	270	94.3	100.0	100.0	0.2	0.0

At these same locations, linear probing predicts the queried state with near-perfect accuracy in the full activation space, yet direct probe-derived steering has near-zero counterfactual recomposition accuracy. The probe identifies which state is present, but its direction rarely makes the model answer according to the source queried state. The complement-space results extend this finding. After subtracting the reference steering subspace component, probing accuracy remains near perfect, showing that the remaining activations retain linearly readable state information outside the reference steering subspace. Directions obtained from this remaining information nevertheless yield near-zero counterfactual recomposition accuracy. Together with the successful reference under the shared intervention conditions, these results distinguish the availability of an effective intervention from its recovery through linear probing. This separation occurs across all eight model–environment conditions, supporting the finding that high readability does not guarantee writability through direct probe-derived directions. Additional reference validation and the same-value source control are reported in Appendices D.1 and E.

4.3 Covariance transformations and combinations of probe directions

We next test whether the read–write gap depends on using individual probe coefficients directly as steering directions. Table 2 compares the constructions in Section 3.5. In the full activation space, covariance-transformed directions improve counterfactual recomposition accuracy over direct probe directions in every model–environment condition. Because both individual-direction constructions

Table 2: Counterfactual recomposition accuracy (%) for all probe-derived steering constructions. Each entry reports full activation space / complement space accuracies. Each accuracy is the mean across canonical-to-canonical and reversed-to-reversed source–target conditions. Single uses the direct probe direction; Cov. single applies the covariance transformation; Multiple combines probe directions; Cov. multiple combines covariance-transformed directions. All conditions use the same eligible pairs as Table 1.

Environment	Model	Single	Cov. single	Multiple	Cov. multiple
ALFWorld	Qwen2.5	0.0/0.0	3.3/0.0	0.5/0.0	9.7/0.0
	Phi-4	0.0/0.0	28.8/0.0	4.7/0.0	55.3/0.0
	Ministral 3	0.0/0.0	2.2/0.0	0.1/0.0	7.3/0.0
	Qwen3	0.8/0.3	27.3/0.0	4.7/0.0	36.2/0.0
WebShop	Qwen2.5	0.2/0.0	11.6/0.0	0.9/0.0	23.3/0.0
	Phi-4	0.0/0.0	30.6/0.0	1.2/0.0	51.1/0.0
	Ministral 3	0.5/0.5	19.1/0.0	0.7/0.0	32.7/0.0
	Qwen3	0.2/0.0	26.5/0.0	3.5/0.0	40.7/0.2

have rank one, this improvement demonstrates that the conversion of predictive coefficients into an intervention direction matters even at fixed rank. Combining multiple untransformed probe directions also improves accuracy, but its gains are consistently smaller than those of the covariance-transformed individual direction. Combining covariance-transformed directions gives the highest accuracy among the probe-derived constructions in every condition, improving on both covariance transformation alone and combinations of untransformed directions.

In the complement space, counterfactual recomposition accuracy remains near zero across all tested constructions and model–environment conditions. The gains observed in the full activation space therefore do not carry over to the complement space, even though the queried state remains nearly perfectly readable there. Allowing the multi-direction combinations to reach the reference rank also leaves the gap intact. The complement-space finding consequently extends beyond the direct use of an individual probe direction to constructions that incorporate activation covariance and multiple state-predictive directions. These two tables establish complementary findings: direction construction affects writability in the full activation space, while high readability outside the reference steering subspace persists without effective steering through any of the tested probe-derived constructions. Cross-presentation results, additional behavioral diagnostics, and uncertainty estimates are provided in Appendix F.

5 Conclusion

We developed a controlled framework for studying the relationship between readability and writability. Counterfactual recomposition provides an evaluation beyond answer and source-plan copying, while a validated reference steering subspace establishes that the required answer change is achievable. Under shared intervention conditions, near-perfect readability in the complement space coexists with near-zero counterfactual recomposition accuracy across all tested probe-derived steering constructions. These findings distinguish the availability of effective steering from its recovery through linear probing, showing why probing accuracy alone is insufficient evidence of writability.

Our framework provides a basis for investigating the conditions under which readable state information can support effective steering. Its task coverage does not extend to long-horizon interaction or free-form generation. The effects of model scale and training regime on the read–write gap also remain to be systematically characterized. Future work could build on this foundation to develop a broader understanding of the relationship between readability and writability and principles for reliable steering.

Reproducibility Statement

The controlled task and intervention constructions are specified in Section 3, with prompt formats and counterfactual group construction detailed in Appendix A. Section 4.1 reports the models, data splits, selection procedures, and reporting conventions. Appendix B provides reference steering subspace

optimization and probe-fitting details, while Appendix C specifies sample eligibility and evaluation criteria. Additional validation and presentation-transfer procedures are documented in Appendix D. The same-value source control is described in Appendix E, and implementation details for covariance transformations and combinations of probe directions, together with uncertainty estimation, are provided in Appendix F.

AI USE STATEMENT

LLM assistance was used to retrieve related work, write code, record experiment results, draw figures, and draft parts of the research paper. The authors specified the research questions and designed the entire experiment. We have checked all AI-assisted work manually, and we take responsibility for the final content of this work, including text, code, claims or artifacts produced with the aid of generative AI.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, et al. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

Raghu Arghal, Fade Chen, Niall Dalton, Evgenii Kortukov, Calum McNamara, Angelos Nalmpantis, Moksh Nirvaan, Gabriele Sarti, and Mario Giulianelli. A behavioural and representational evaluation of goal-directedness in language model agents, 2026. URL https://arxiv.org/abs/2602.08964.

Aryaman Arora, Dan Jurafsky, and Christopher Potts. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14638–14663, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.785. URL https://aclanthology.org/2024.acl-long.785/.

Yuntai Bao, Xuhong Zhang, Tianyu Du, Xinkui Zhao, Zhengwen Feng, Hao Peng, and Jianwei Yin. Probing the geometry of truth: Consistency and generalization of truth directions in LLMs across logical transformations and question answering tasks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 682–700, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.38. URL https://aclanthology.org/2025.findings-acl.38/.

Marc E. Canby, Adam Davies, Chirag Rastogi, and Julia Hockenmaier. How reliable are causal probing interventions? In *Proceedings of IJCNLP-AACL*, pp. 857–878. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.ijcnlp-long.47. URL https://aclanthology.org/2025.ijcnlp-long.47/.

Alicja Dobrzeniecka, Antske Fokkens, and Pia Sommerauer. Improving causal interventions in amnesic probing with mean projection or LEACE. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 12981–12993. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.674. URL https://aclanthology.org/2025.findings-acl.674/.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations.

In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating interpretability methods on disentangling language model representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8669–8687, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.470. URL https://aclanthology.org/2024.acl-long.470/.

Sadegh Khorasani, Saber Salehkaleybar, Negar Kiyavash, and Matthias Grossglauser. Inference time causal probing in LLMs, 2026. URL https://arxiv.org/abs/2605.07631.

Michael A. Lepori, Thomas Serre, and Ellie Pavlick. Uncovering intermediate variables in transformers using circuit probing. In *First Conference on Language Modeling*. OpenReview.net, 2024. URL https://openreview.net/forum?id=gUNeyiLNxr.

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*, volume 36, pp. 41451–41530, 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, et al. Ministral 3, 2026. URL https://arxiv.org/abs/2601.08584.

Gabriele Maraia, Leonardo Ranaldi, Marco Valentino, and Fabio Massimo Zanzotto. Can activation steering generalize across languages? a study on syllogistic reasoning in language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2739–2753, Rabat, Morocco, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.125. URL https://aclanthology.org/2026.eacl-long.125/.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Qwen Team. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of EACL*, pp. 3363–3377. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295/.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of ACL*, pp. 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Aryan Sharma, Cutter Dawes, and Shivam Raval. Dissociating decodability and causal use in bracket-sequence transformers, 2026. URL https://arxiv.org/abs/2604.22128.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=0IOX0YcCdTn.

Jan Sobotka, Mustafa O. Karabag, and Ufuk Topcu. Why do LLMs struggle in strategic play? broken links between observations, beliefs, and actions, 2026. URL https://arxiv.org/abs/2605.00226.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12964–12974, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.634. URL https://aclanthology.org/2025.acl-long.634/.

Devesh Tiwari, Camille Davis, Shivank Sinha, Talia Weaver, Aditya Shah, and Maheep Chaudhary. Decodability is not causality: Dissociating probe readouts from behavioral drivers via SAE decomposition, 2026. URL https://arxiv.org/abs/2609.18080.

Sihan Wang, Jiayi Zhao, Qingyan Cao, Hongbo Yao, and Lin Shu. FishBack: Pullback fisher geometry for optimal activation steering in transformers, 2026. URL https://arxiv.org/abs/2605.17231.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. Interpretability at scale: Identifying causal mechanisms in alpaca. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78205–78226, 2023. doi: 10.52202/075280-3418. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f6a8b109d4d4fd64c75e94aaf85d9697-Abstract-Conference.html.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, volume 35, pp. 20744–20757, 2022. doi: 10.52202/068431-1508. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/82ad13ec01f9fe44c01cb91814fd7b8c-Abstract-Conference.html.

A Controlled Task Construction

Prompt format. Each prompt has four blocks: environment name, available actions, a two-row rule, and a four-row answer mapping. A representative schematic is:

```
Controlled decision task. Use only the listed actions.

[AVAILABLE ACTIONS]
- Action kind "access" -- action: open cabinet 1
- Action kind "submit" -- action: put object 1 ...

[DECISION RULE]
- If at least one listed action has action kind "access",
  choose PLAN_2.
- If no listed action has action kind "access",
  choose PLAN_4.

[OUTPUT CODEBOOK]
- PLAN_1 -> 7
- PLAN_2 -> 1
- PLAN_3 -> 9
- PLAN_4 -> 4

Return exactly one code and no explanation.
```

Table 3: Atomic transformations used to construct and validate a counterfactual group.

Transformation	Changed field	Required semantic relation
Queried-state flip	one action row	plan and answer change
Unqueried-state flip	one action row	plan and answer unchanged
Action replacement	one action string	state bits unchanged
Rule-query flip	queried role	plan and answer change
Selected-branch replacement	one branch plan	plan and answer change
Unselected-branch replacement	one branch plan	plan unchanged
Selected-answer swap	two answer-mapping entries	plan unchanged, answer changes
Unselected-answer swap	two answer-mapping entries	answer unchanged
Rule-order reversal	row order only	all semantics unchanged

Table 4: Selected intervention layers at the final prompt position and reference steering subspace ranks. A layer pair gives the intervention layer and the selected later layer used for the blocking audit. Indices follow the local model implementation; intervention depth is its index divided by the model's transformer-layer count.

Environment	Model	Layers	Intervention depth	Rank
ALFWorld	Qwen2.5	34→35	0.708	8
	Phi-4	23→24	0.575	8
	Ministral 3	27→28	0.675	9
	Qwen3	27→28	0.675	8
WebShop	Qwen2.5	34→35	0.708	9
	Phi-4	23→24	0.575	8
	Ministral 3	23→24	0.575	8
	Qwen3	27→28	0.675	8

The `OUTPUT CODEBOOK` block specifies the answer mapping. The reversed presentation swaps only the two decision-rule rows. The generator checks that every other structured field remains byte-identical. Answers are drawn without replacement from the ten digits. Model preflight checks that all ten digits are distinct single tokens in the answer context.

Atomic transformations. Table 3 lists the controlled input changes used for task construction, probe fitting, and supplementary validation.

Adding or removing an action changes exactly one action row and one state bit. Action replacement changes one row at token edit distance one and must retain the same action role. Answer-mapping changes swap the answers assigned to two plan labels.

Counterfactual group construction. For an opposite-value pair, the target is the base task. The source combines a queried-state flip, a replacement of the branch plan that was unselected in the target but becomes selected in the source, and an answer-mapping swap. The generator checks the plan and answer separation conditions in Eqs. 3 and 4. The direct counterfactual changes only the queried-state action in the target. We score the three distinct answers: the original target answer A, the counterfactual answer A^{cf}, and the source-plan answer $C(Q')$. The source answer $C'(Q')$ equals the original target answer. The counterfactual answer is the sole success criterion.

B Implementation and Optimization Details

Intervention layer selection. All interventions act on the residual stream at the shared final prompt position. In each layer pair in Table 4, Eq. 5 is applied at the first layer, and target restoration at the second tests whether the transferred effect is removed. The search constrains the later layer only to

follow the intervention layer; the adjacent pairs in Table 4 are outcomes of this search rather than an imposed adjacency constraint.

The layer scan uses the deterministically selected 24-of-32 selection groups at the final prompt position; the clean source/target behavior check retains 23–24 groups in the eight model–environment conditions. It screens residual-stream layer pairs using full source replacement and target restoration at the later layer. Effects at the intervention and later layers are normalized by the corresponding direct-counterfactual logit-gap effect; mediation is the fraction of the intervention-layer effect removed by target restoration at the later layer. Source and output controls use the same normalization. The prespecified screen requires intervention-layer effect at least 0.20, mediation of that effect at least 0.20, later-layer effect at least 0.30, later-layer role accuracy at least 0.50, correct- over-shifted pairing gain at least 0.10, and output-control effect at most 0.35. Candidates are ranked lexicographically by whether all margins pass, the number of passing margins, the worst margin, pairing gain, later-layer accuracy, smaller output effect, and the aggregate path score.

Reference steering subspace initialization and optimization. Each paired training item supplies forward and reverse activation differences Δ^+, Δ^- and logit-gap gradients g^+, g^- at the selected intervention layer. Let $z_a(x)$ denote the model's logit for answer a on unmodified input x. We compute

$$\begin{aligned} \Delta^+ &= h' - h, & g^+ &= \nabla_h (z_{A^{\mathrm{cf}}}(x) - z_A(x)), \\ \Delta^- &= h - h^{\mathrm{cf}}, & g^- &= \nabla_{h^{\mathrm{cf}}} \left(z_A(x^{\mathrm{cf}}) - z_{A^{\mathrm{cf}}}(x^{\mathrm{cf}}) \right). \end{aligned} \tag{17}$$

The $+$ signal favors counterfactual recomposition; the $-$ signal favors restoration of the original target answer from the direct counterfactual. Pooling both signals across training items and the two presentations gives N difference–gradient pairs, indexed as (Δ_i, g_i). The initializer is the leading eigenspace of the symmetric matrix

$$H = \frac{1}{2N} \sum_{i=1}^{N} \left(\Delta_i g_i^\top + g_i \Delta_i^\top \right). \tag{18}$$

The initializer supplies the ordered basis B for the learned boundary in Eq. 8. Training then minimizes Eq. 9, using the forward and reverse replacements above. The normalized replacement-energy penalty is defined in Eq. 10.

Fitting begins at width $m = \min(16, d, 2N)$ and expands by 16, up to $\min(d, 2N)$, only when all current coordinates survive the hard threshold and the smallest computed initializer eigenvalue exceeds 10^{-6} times the leading eigenvalue. Coordinates with $a_i \geq 0.5$ are retained and the resulting hard basis is refined. Three fixed restarts are compared using the selection rule below. The selected ranks are reported in Table 4 and fixed before final evaluation. One basis is fitted per model–environment condition and shared across queried states, rules, plans, answer mappings, and both presentations.

Model forward passes use bfloat16; basis optimization, gradients, and summary statistics use float32 where required. The initial computational width and increment are 16; power iteration uses six iterations with oversampling eight. Soft-boundary optimization uses 80 steps followed by 40 hard-refinement steps, learning rate 0.005, batch size four, replacement-energy weight $\lambda_E = 0.02$, and effective-rank weight $\lambda_r = 0.01$. Temperature τ decreases linearly from 2.0 to 0.1. The three restart seeds are 20261201, 20261202, and 20261203. The energy penalty in Eq. 10 uses minibatch averages and $\epsilon = 10^{-8}$, with soft P during boundary optimization and $P = UU^\top$ during hard refinement. The hard refinement removes the effective-rank term but retains the two cross-entropies and normalized-energy penalty.

Candidates are ranked lexicographically on selection data. First, we maximize the smaller, across the two presentations, sum of counterfactual recomposition accuracy, restoration accuracy, and semantic selectivity. Ties are broken by higher worst-presentation values of those three metrics in that order, then by lower worst-presentation mean Jensen–Shannon divergence between intervened-target and direct-counterfactual answer distributions, and finally by lower rank. Selection uses the two 24-group presentation sets; the layer, rank, restart, and basis are not updated during review or final evaluation.

Linear probe fitting. Ridge classifiers use at most 2,000 iterations, tolerance 10^{-4}, and regularization grid $\{0.1, 1, 10, 100\}$. Features are standardized using training-set statistics, and regularization

is selected on the 32 selection groups. Coefficients are converted to steering directions using Eq. 12, with the additional complement-space projection for probes fitted in that space.

C Evaluation Protocol and Supplementary Metrics

Data partitions and reporting unit. The generator creates 160 training, 32 selection, 48 review, and 416 final groups per environment. Deterministic group-ID sorting selects the first 24 selection groups for intervention-layer, reference-subspace, rank, and restart selection; all 32 are available only to choose ridge regularization. The review split assesses the selected layer and basis without reselection, and the primary comparison evaluates the final split once. An independently generated atomic group is the statistical unit within each fixed model–environment condition.

Sample eligibility. For each source–target pair and presentation, the unmodified model is scored over the four digits in the relevant answer mapping. A group enters intervention evaluation only when the model selects the correct answer for the source, target, and direct counterfactual under both rule orders. This eligible intersection is fixed before any intervention metric is computed. Because eligibility depends on model behavior, each model–environment condition evaluates a different subset of the same 416 raw final groups; the resulting counts range from 270 to 415.

For reference steering subspace fitting and evaluation, the clean-answer filtering leaves 103–160 training groups, 16–24 of the 24 selection groups, and 28–48 review groups across the eight model–environment conditions. For probing in Tables 1 and 7, all final groups are evaluated without this filter: the canonical set includes base prompts and the eight atomic transformations other than rule-order reversal, whereas the reversed set contains only base prompts with the rule rows reordered.

Supplementary metrics and decision criteria. The main comparison reports counterfactual recomposition accuracy as defined in Eq. 6. The following supplementary diagnostics evaluate source–target pairing, competing answers, reverse restoration, downstream blocking, and matched-rank random subspaces. Their constructions and results are reported for the reference steering subspace in Appendix D.1 and for probe-derived steering in Appendix F.

For each intervention output, role accuracy is the fraction of predictions equal to the answer associated with a semantic role: factual target, counterfactual queried state, source plan under the target answer mapping, source answer, or direct copy. Rivals whose answer is never distinguishable from the counterfactual answer are excluded. Semantic selectivity subtracts the highest accuracy among the remaining rival roles from counterfactual recomposition accuracy. The pairing gap subtracts counterfactual accuracy after the circular source shift from accuracy with the correct source–target pairing.

Let $\gamma = z_{A^{\mathrm{cf}}} - z_A$ be the counterfactual-versus-factual logit gap. Write γ_t, γ_c, γ_{steer}, and γ_{block} for this gap on the clean target, clean direct counterfactual, intervened target, and intervened target followed by target restoration at the blocking layer, respectively. Their mean effects relative to the clean target are

$$\begin{aligned} e_{\mathrm{CF}} &= \mathbb{E}[\gamma_c - \gamma_t], \\ e_{\mathrm{steer}} &= \mathbb{E}[\gamma_{\mathrm{steer}} - \gamma_t], \\ e_{\mathrm{block}} &= \mathbb{E}[\gamma_{\mathrm{block}} - \gamma_t]. \end{aligned} \tag{19}$$

The expectation averages eligible pairs within each presentation condition. The effect ratio is $e_{\mathrm{steer}}/e_{\mathrm{CF}}$, and the blocked fraction is $(e_{\mathrm{steer}} - e_{\mathrm{block}})/e_{\mathrm{steer}}$. Restoration accuracy is factual-answer accuracy when a direct counterfactual prompt receives the target component at the intervention layer. Ratios are computed from the condition-level mean effects above.

A reference qualifies as a positive control for writability only if it has at least 48 eligible final groups, counterfactual accuracy at least 0.60, semantic selectivity at least 0.20, pairing gain at least 0.15, effect ratio at least 0.20, blocked fraction at least 0.20, restoration accuracy at least 0.60, and a gain of at least 0.10 over the matched-rank random 95th percentile.

These thresholds were specified before the primary final evaluation. They are operational safeguards: 0.60 requires majority-level transfer and restoration, while the margin criteria require effects separated

from semantic rivals, incorrect pairing, and random spaces. We report continuous values alongside every primary pass/fail verdict.

D Supplementary Validation and Presentation Transfer

These experiments supplement the comparison of readability and writability in Section 4.2. We report held-out checks of the selected intervention locations, additional validation of the reference steering subspace, and transfer across rule-row orders. Each evaluation uses the previously selected intervention quantities and its designated review or final split.

D.1 Intervention validation

Held-out intervention-location diagnostics. The review split evaluates the queried-state localization pairs used in Appendix B, keeping the selected layers fixed. An attention/MLP component pair is selected within these layers on selection data. Component mediation measures the fraction of its source-replacement effect removed by restoring the later component to its target activation. Paired presentation rewrites change one feature at a time: action-field or action-list order, rule wording or row order, answer-mapping notation or row order, or output-instruction wording. For each rewrite, we compare the normalized intervention-layer, blocking, and later-layer effects on the same eligible examples and report the largest absolute difference. Decision-preserving controls change an unqueried state, an unselected rule branch, or unselected answer mappings; their largest prediction-change rate is reported separately. Each paired rewrite requires at least 12 eligible examples; the diagnostic cutoffs are specified in Table 5.

Table 5: Held-out review of the queried-state intervention locations. Component mediation is the fraction of the component intervention effect removed by restoring the later component. Rewrite gap is the largest absolute change in normalized residual intervention effects across paired presentation rewrites. Control change is the largest prediction-change rate on decision-preserving controls. Values are percentages; n_{rewrite} is the smallest paired sample count. Diagnostic cutoffs are 15%, 20%, and 10%, respectively.

Environment	Model	n_{rewrite}	Component mediation	Rewrite gap	Control change
ALFWorld	Qwen2.5	35	32.8	4.4	0.0
	Phi-4	35	22.1	3.4	0.0
	Ministral 3	35	-275.0	5.8	0.0
	Qwen3	34	21.6	3.7	2.2
WebShop	Qwen2.5	32	8.4	2.5	0.0
	Phi-4	35	21.5	1.5	0.0
	Ministral 3	32	285.7	4.5	0.0
	Qwen3	34	20.2	6.6	0.0

Table 5 shows that all conditions meet the paired-sample requirement and pass the rewrite and decision-preserving checks. The selected residual intervention effects remain stable under the tested presentation changes, while decision-preserving input changes rarely alter predictions. Component mediation varies substantially across models and environments; this signed-effect ratio is not bounded like an accuracy. These review diagnostics evaluate the selected locations; the learned reference steering subspaces are evaluated separately below.

Reference steering subspace validation. A circular shift pairs each target with the next source while preserving the marginal activation distribution. Matched-rank random orthonormal subspaces test whether dimension alone accounts for the learned subspace's effect. Semantic selectivity compares the counterfactual answer with distinguishable competing answers. Reverse restoration replaces the direct counterfactual activation's subspace component with the target component; downstream blocking restores the target activation at the selected later layer after a forward intervention. We evaluate 32 random orthonormal subspaces at ranks one and r_{ref}. The random-subspace criterion uses their 95th-percentile accuracy; descriptive summaries report the observed maximum. Every steering construction uses the projected source–target difference in Eq. 5 with coefficient one. Appendix C defines the effect ratios and aggregate validation criteria.

The reference steering subspaces pass all of these prespecified criteria in every model–environment condition. Table 6 reports the less favorable presentation for each diagnostic, retaining the original thresholds and continuous scores.

Table 6: Supplementary validation of the reference steering subspace. Each column reports the lower value across canonical-to-canonical and reversed-to-reversed steering. Pair gap uses a circular source shift; effect ratio normalizes by the direct-counterfactual effect; blocked is the fraction removed by later target restoration; restore is reverse replacement accuracy; random gap uses the matched-rank 95th percentile. Required minima are 20, 15, 20, 20, 60, and 10%, respectively; every condition passes. Signed-effect ratios may exceed 100%.

Environment	Model	Select.	Pair gap	Effect ratio	Blocked	Restore	Random gap
ALFWorld	Qwen2.5	89.66	55.67	73.13	99.77	97.29	93.84
	Phi-4	95.42	53.98	79.89	99.96	100.00	97.11
	Ministral 3	74.34	53.97	69.77	100.20	97.88	87.04
	Qwen3	89.07	54.34	83.43	99.97	95.18	93.57
WebShop	Qwen2.5	82.11	50.88	72.54	100.06	95.44	90.18
	Phi-4	97.76	49.50	79.97	99.95	99.50	98.76
	Ministral 3	88.82	49.01	78.57	99.83	99.34	94.08
	Qwen3	85.19	40.00	87.46	99.83	97.78	90.00

The audit measures complementary aspects of the intervention. Correct pairing supports dependence on the paired source value; semantic selectivity measures preference for the target-side counterfactual role; and reverse replacement recovers the factual answer. Restoring the full target residual at the later layer resets the downstream computation. Agreement across these tests supports a counterfactual recomposition account beyond output accuracy alone. The matched-rank comparison shows that the learned reference steering subspaces outperform the sampled random subspaces of the same dimension.

D.2 Cross-presentation steering and probing within the reference steering subspace

The canonical presentation lists the $S_{j(G)} = 1$ rule row first; the reversed presentation swaps only the two rule rows. The action list, the condition-to-plan assignments, the answer mapping, and the correct answer remain unchanged. This construction tests sensitivity to rule-row order without changing the task's computation.

Cross-presentation evaluation uses a canonical source with a reversed target, or a reversed source with a canonical target, while retaining the same sample identities, layer, and steering basis. In addition, a separate linear ridge probe is fitted to activations restricted to the reference steering subspace to measure queried-state readability within it.

Table 7: Reference steering across rule-row orders and probing within the reference steering subspace. C and R denote canonical and reversed order; arrows run from source to target. Steering columns report counterfactual recomposition accuracy (%). Ref. probe reports the lower queried-state balanced accuracy across the two rule-row orders (%). Eligible n applies to steering.

Environment	Model	Eligible n	C→C	R→R	C→R	R→C	Ref. probe
ALFWorld	Qwen2.5	406	95.8	93.8	82.0	74.9	71.7
	Phi-4	415	99.3	97.1	98.8	85.8	100.0
	Ministral 3	378	87.0	89.9	74.1	85.2	100.0
	Qwen3	311	93.6	96.1	93.2	85.5	98.6
WebShop	Qwen2.5	285	90.2	90.5	73.3	76.5	96.0
	Phi-4	402	98.8	99.0	92.8	88.6	99.3
	Ministral 3	304	94.4	98.0	90.5	94.4	99.8
	Qwen3	270	97.8	90.7	88.1	92.2	99.8

Table 7 reports the four source–target rule-order combinations separately and adds probing accuracy within the reference steering subspace. Averaging C→C and R→R recovers the reference steering results in Table 1. Accuracy depends on the source–target rule-row combination. Qwen2.5 on

ALFWorld shows the largest cross-presentation drop and lower probing accuracy within the reference steering subspace than the other model–environment conditions. Thus, the reference supports counterfactual recomposition across rule-row orders even in a condition where its queried state is less reliably recovered by linear probing.

The same rule-row transformation is also applied to the same-value source control, whose construction and results are reported in Appendix E.

E SAME-VALUE SOURCE CONTROL

This control supplements the opposite-value results in Section 4.2 by testing whether the reference steering subspace preserves the target answer when the source queried state is unchanged. An intervention that always switches the target rule branch would instead change the answer. The same-value source $x' = (S', G', C')$ is constructed for the same target and queried coordinate. Writing $j = j(G)$, the generator enforces

$$j(G') = j, \qquad S'_j = S_j, \qquad Q' \neq Q, \qquad C'(Q') = A, \tag{20}$$

where $Q' = G'(S')$. Relative to the target, the source changes one unqueried state coordinate, its selected branch plan, and a transposition involving its selected answer. The target answer mapping assigns Q' an answer distinct from A. Transferring the already-matching queried value should therefore retain the original target answer A. Copying the source plan under the target answer mapping or switching the target branch changes the answer. Copying the source answer alone is indistinguishable from factual retention here and is addressed by the opposite-value construction.

The direct source-rule control applies the source's selected-branch replacement to the target state and target answer mapping, for which the correct answer is $C(Q')$. Eligibility requires correct clean predictions for the source, target, and direct source-rule control under both presentations. This checks that the model can distinguish the source-plan alternative even though the queried state is unchanged. The model-specific eligible-pair set $\mathcal{E}_F^{=}$ covers 295–416 of the 416 final group identities. Using x' as the source in Eq. 5, factual retention under the frozen reference is

$$\mathrm{Ret}^{v_s \to v_t}(U) = \Pr\left[F\left(x^{v_t}; \mathrm{do}(h = \widetilde{h})\right) = A \mid (x', x) \in \mathcal{E}_F^{=}\right]. \tag{21}$$

The protocol and thresholds were fixed before final evaluation. The control reuses the final group identities and frozen intervention layer, reference basis, rank, and replacement coefficient one; it does not refit or reselect an intervention quantity. Each same-presentation condition must contain at least 48 eligible groups, retain at least 90% of factual predictions, change at most 10% of clean predictions, and select the source plan under the target answer mapping on at most 10%. Both same-presentation conditions must pass in every model–environment condition. Cross-presentation conditions are descriptive. We compute 95% intervals using 10,000 group-bootstrap draws with seed 20260803.

Table 8 reports the less favorable same-presentation retention rate and the larger error rates.

Table 8: Same-value source control with the frozen reference steering subspace. Retain is the less favorable target-answer retention rate; Change and Source plan are the larger error rates across the two same-presentation conditions. Eligible sample counts use the separate same-value eligibility rule.

Environment	Model	Eligible n	Retain (%)	Change (%)	Source plan (%)
ALFWorld	Qwen2.5	416	99.5	0.5	0.0
	Phi-4	410	95.4	4.6	3.9
	Ministral 3	392	98.2	1.8	0.3
	Qwen3	317	99.7	0.3	0.0
WebShop	Qwen2.5	321	95.6	4.4	0.9
	Phi-4	398	99.0	1.0	0.5
	Ministral 3	357	99.2	0.8	0.3
	Qwen3	295	92.2	7.8	1.0

The reference retains the target answer at a high rate in every model–environment condition, while selection of the source plan under the target answer mapping remains rare. All conditions pass the

prespecified same-presentation criteria. Together with the opposite-value result, this excludes an indiscriminate branch switch for the tested source construction. Under the rule-row transformation defined in Appendix D.2, Table 9 reports the two cross-presentation directions separately. Retention remains high in most conditions, but reversed-to-canonical steering reduces retention for both Qwen models on WebShop. The same-value result therefore depends on whether source and target use the same rule-row order.

Table 9: Same-value source control across rule-row orders. Each entry reports C→R / R→C, where C and R denote canonical and reversed order. Retain is target-answer retention; Change is prediction change; Source plan is selection of the source plan under the target answer mapping. These descriptive results use the same eligible groups and reference steering subspace as Table 8.

Environment	Model	Eligible n	Retain (%)	Change (%)	Source plan (%)
ALFWorld	Qwen2.5	416	98.1/92.3	1.9/7.7	0.0/0.0
	Phi-4	410	94.1/97.3	5.9/2.7	4.1/1.2
	Ministral 3	392	95.7/97.7	4.3/2.3	0.0/0.5
	Qwen3	317	99.1/98.7	0.9/1.3	0.0/0.0
WebShop	Qwen2.5	321	95.0/83.2	5.0/16.8	0.9/5.6
	Phi-4	398	95.5/98.7	4.5/1.3	2.3/0.5
	Ministral 3	357	99.2/99.2	0.8/0.8	0.3/0.0
	Qwen3	295	95.6/85.8	4.4/14.2	0.3/1.4

F Supplementary Analysis of Probe-Derived Steering

The main construction and its counterfactual recomposition results are reported in Sections 3.5 and 4.3. Here we record implementation details and additional diagnostics.

F.1 Implementation details

For a primary probe coefficient w, the covariance-transformed direction in Eq. 13 is computed as a covariance–vector product on centered training activations without forming or inverting a dense covariance matrix.

The candidate pool contains 13 probe coefficients: four state-coordinate probes under canonical rule-row order, four under reversed rule-row order, four queried-state probes conditioned on the queried coordinate, and one pooled queried-state probe. Canonical state-coordinate probes pool the base prompt and the eight semantic transformations in Table 3 other than rule-order reversal. Reversed probes use base semantics with only the two rule rows reordered, while query-conditioned and pooled probes use base semantics under both orders. Coefficients are fitted on training data and regularization is selected on selection data. Both the direct and covariance-transformed combinations normalize each column before singular value decomposition, as defined in Eqs. 14 and 16. The numerical rank counts singular values exceeding 10^{-6} times the largest singular value; the retained rank is capped by the reference rank. For complement-space combinations, the retained basis is additionally projected by $I - UU^\top$ and orthonormalized, with its actual numerical rank recorded.

F.2 Supplementary evaluation

The supplementary evaluation reuses the selected intervention and blocking layers, eligible final groups, replacement operation, validation criteria, and random controls. Final groups are used only for evaluation. In addition to the same-presentation accuracies in Table 2, we evaluate cross-presentation steering using the rule-row transformation in Appendix D.2 and the behavioral diagnostics in Appendix D.1, with criteria defined in Appendix C. For each construction, both full activation space and complement space results are reported; summary comparisons use the space with the higher average same-presentation accuracy. Random controls from the primary analysis are reused when ranks match. Uncertainty intervals use 10,000 paired bootstrap draws over atomic groups at fixed model–environment conditions (95% confidence; seed 20260801). Figure 2 displays the main accuracy comparison with uncertainty estimates. Columns separate environments and rows identify models.

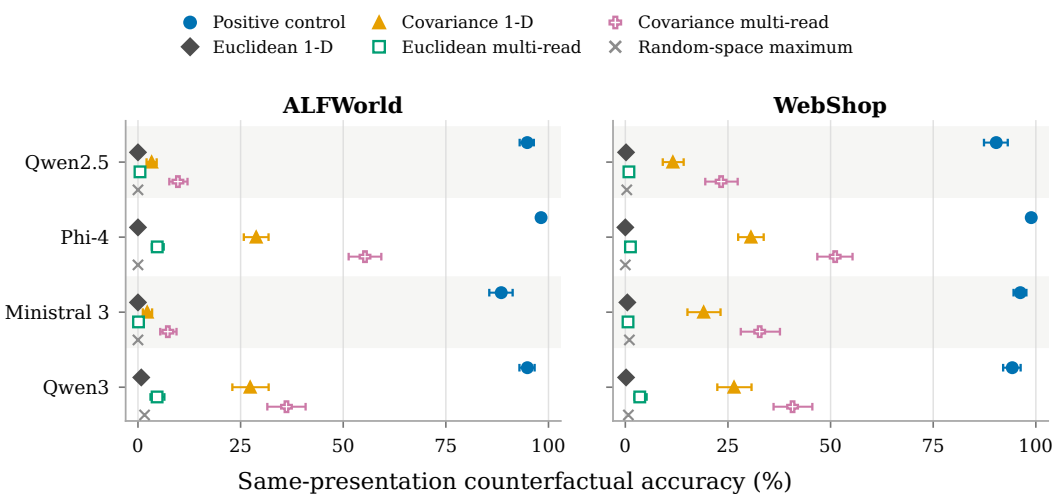


Figure 2: Counterfactual recomposition accuracy on eligible final groups. Each probe-derived construction uses the full activation space or complement space variant with higher mean same-presentation accuracy. Points average canonical-to-canonical and reversed-to-reversed steering; whiskers show 95% group-bootstrap intervals. Random markers show the maximum across evaluated ranks and same-presentation conditions.

Table 10 reports behavioral diagnostics for Phi-4 with combinations of covariance-transformed directions in the full activation space, the highest-accuracy probe-derived construction in the main comparison. Counterfactual recomposition and semantic selectivity remain weaker than for reference steering, and reverse restoration is less reliable under reversed rule-row order in both environments. Panel (a) reports counterfactual recomposition accuracy, semantic selectivity, and the correct-over-shifted pairing gap. Panel (b) reports the effect ratio, blocked fraction, reverse restoration accuracy, and gain over the matched-rank random-subspace 95th percentile. Effect ratio and blocked fraction are defined in Appendix C.

Table 10: Behavioral diagnostics for probe-derived steering using Cov. multiple in the full activation space of Phi-4. Values are percentages except n; headers show the prespecified intervention thresholds. Pair is the correct-over-shifted pairing gain, Effect is the direct-counterfactual-normalized intervention effect, Blocked is the fraction removed by later target restoration, Restore is reverse replacement accuracy, and Random is the gain over the matched-rank random-subspace 95th percentile.

(a) Counterfactual recomposition and selectivity

Environment	Order	$n \geq 48$	CF acc. ≥ 60	Select. ≥ 20	Pair ≥ 15
ALFWorld	Canonical	415	56.9	13.7	35.9
	Reversed	415	53.7	7.7	32.0
WebShop	Canonical	402	56.7	13.4	30.1
	Reversed	402	45.5	-9.0	23.6

(b) Intervention effects and controls

Environment	Order	Effect ≥ 20	Blocked ≥ 20	Restore ≥ 60	Random ≥ 10
ALFWorld	Canonical	52.2	99.9	62.4	56.9
	Reversed	49.2	100.0	54.2	53.7
WebShop	Canonical	55.5	100.0	68.7	56.7
	Reversed	46.2	99.9	51.7	45.5

Table 11 reports each construction separately in the full activation space and complement space, denoted F/C. Same and Cross average the corresponding two presentation conditions. Gate reports whether both same-presentation contexts satisfy every transfer, semantic-selectivity, pairing, restoration, blocking, and matched-rank random criterion. Δ is reference steering accuracy minus the higher accuracy across the two spaces, with a paired 95% group-bootstrap interval in percentage points. The reference maintains higher counterfactual recomposition accuracy than every probe-derived construction, including the full-space constructions that improve accuracy in Section 4.3.

Table 11: Supplementary evaluation of probe-derived steering. F/C denotes full activation space / complement space. Same and Cross report mean accuracy (%) over C→C and R→R, and C→R and R→C, respectively; C and R denote canonical and reversed rule-row order. Gate reports whether both same-presentation conditions meet all criteria in Appendix C. Δ is reference steering accuracy minus the higher Same accuracy across the two spaces, with a paired 95% group-bootstrap interval in percentage points.

Env.	Model	Construction	Rank F/C	Same F/C	Cross F/C	Gate F/C	Δ [95% CI]
ALFWorld	Qwen2.5	Single	1/1	0.0/0.0	0.0/0.0	N/N	+94.8 [+93.1, +96.6]
		Cov. single	1/1	3.3/0.0	1.8/0.0	N/N	+91.5 [+89.4, +93.5]
		Multiple	8/8	0.5/0.0	0.6/0.0	N/N	+94.3 [+92.5, +96.1]
		Cov. multiple	8/8	9.7/0.0	4.8/0.0	N/N	+85.1 [+82.4, +87.7]
	Phi-4	Single	1/1	0.0/0.0	0.0/0.0	N/N	+98.2 [+97.2, +99.0]
		Cov. single	1/1	28.8/0.0	21.6/0.0	N/N	+69.4 [+66.1, +72.5]
		Multiple	8/8	4.7/0.0	4.8/0.0	N/N	+93.5 [+91.7, +95.1]
		Cov. multiple	8/8	55.3/0.0	34.9/0.0	N/N	+42.9 [+38.9, +46.9]
	Ministral 3	Single	1/1	0.0/0.0	0.0/0.0	N/N	+88.5 [+85.6, +91.3]
		Cov. single	1/1	2.2/0.0	1.6/0.0	N/N	+86.2 [+83.2, +89.2]
		Multiple	9/9	0.1/0.0	0.1/0.0	N/N	+88.4 [+85.3, +91.1]
		Cov. multiple	9/9	7.3/0.0	3.8/0.0	N/N	+81.2 [+77.9, +84.4]
	Qwen3	Single	1/1	0.8/0.3	0.6/0.2	N/N	+94.1 [+92.1, +95.8]
		Cov. single	1/1	27.3/0.0	23.8/0.0	N/N	+67.5 [+63.0, +71.9]
		Multiple	8/8	4.7/0.0	5.0/0.0	N/N	+90.2 [+87.8, +92.4]
		Cov. multiple	8/8	36.2/0.0	29.1/0.0	N/N	+58.7 [+54.2, +63.2]
WebShop	Qwen2.5	Single	1/1	0.2/0.0	0.2/0.0	N/N	+90.2 [+87.2, +93.0]
		Cov. single	1/1	11.6/0.0	7.2/0.0	N/N	+78.8 [+75.3, +82.3]
		Multiple	9/9	0.9/0.0	0.9/0.0	N/N	+89.5 [+86.3, +92.5]
		Cov. multiple	9/9	23.3/0.0	15.8/0.0	N/N	+67.0 [+62.5, +71.6]
	Phi-4	Single	1/1	0.0/0.0	0.0/0.0	N/N	+98.9 [+98.0, +99.6]
		Cov. single	1/1	30.6/0.0	24.6/0.0	N/N	+68.3 [+65.0, +71.5]
		Multiple	8/8	1.2/0.0	1.2/0.0	N/N	+97.6 [+96.4, +98.8]
		Cov. multiple	8/8	51.1/0.0	41.7/0.0	N/N	+47.8 [+43.4, +52.2]
	Ministral 3	Single	1/1	0.5/0.5	0.5/0.2	N/N	+95.7 [+93.9, +97.4]
		Cov. single	1/1	19.1/0.0	19.4/0.0	N/N	+77.1 [+72.9, +81.2]
		Multiple	8/8	0.7/0.0	0.8/0.0	N/N	+95.6 [+93.8, +97.2]
		Cov. multiple	8/8	32.7/0.0	31.2/0.0	N/N	+63.5 [+58.6, +68.3]
	Qwen3	Single	1/1	0.2/0.0	0.2/0.0	N/N	+94.1 [+91.7, +96.1]
		Cov. single	1/1	26.5/0.0	23.7/0.0	N/N	+67.8 [+63.1, +72.2]
		Multiple	8/8	3.5/0.0	3.7/0.0	N/N	+90.7 [+88.0, +93.3]
		Cov. multiple	8/8	40.7/0.2	34.6/0.2	N/N	+53.5 [+48.7, +58.3]

Table 12 reports probing accuracy within the direction or subspace used for each summary comparison. The queried state remains highly readable, whereas some constructions predict the four fixed state coordinates less accurately. Each entry reports balanced accuracy for the queried state followed by the minimum balanced accuracy over the four fixed state coordinates, taking the less favorable rule-row order for each quantity.

Table 12: Probing within each probe-derived steering direction or subspace. Each construction uses the full activation space or complement space variant with higher mean same-presentation steering accuracy. Entries give queried-state balanced accuracy / the minimum balanced accuracy over the four fixed state coordinates (%), taking the lower value across canonical and reversed rule-row orders for each quantity.

		Minimum balanced accuracy (%)			
Environment	Model	Single	Cov. single	Multiple	Cov. multiple
ALFWorld	Qwen2.5	100.0/50.0	98.9/50.0	100.0/81.3	97.2/47.9
	Phi-4	100.0/50.0	100.0/50.0	100.0/83.8	100.0/67.4
	Ministral 3	100.0/50.0	99.6/50.0	100.0/89.4	100.0/50.0
	Qwen3	100.0/50.0	99.4/50.0	100.0/87.6	99.1/50.0
WebShop	Qwen2.5	100.0/50.0	88.5/50.0	100.0/89.6	96.9/58.5
	Phi-4	100.0/50.0	99.8/50.0	100.0/68.5	99.5/50.0
	Ministral 3	100.0/50.0	100.0/50.0	100.0/88.1	99.8/65.3
	Qwen3	100.0/50.0	100.0/50.0	100.0/85.5	99.3/61.7