BEHAVIORAL COMPOSITIONALITY OF SPARSE AUTOENCODER REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse autoencoders (SAEs) represent model activations with sparse, linearly decoded features, but the downstream effects of those features need not compose additively. We introduce SAE Additivity Residuals (SAR), which quantify the behavioral variation left unexplained above a chosen interaction order. Under independent feature subset sampling, population SAR is exactly the normalized energy of interactions above that order and bounds coalition ordering error for the optimal low order approximation. To test whether compositionality under subset interventions depends on coordinate choice beyond the selected subspace and downstream continuation, we compare SAE coordinates with sum preserving rotations of the same subspace and find systematically lower first order residuals. Yet the SAE coordinates are not fully additive: pairwise terms improve held out coalition ordering in each setting, raising average rank fidelity from 0.9782 to 0.9959, and outperform parameter matched models using randomly selected third order terms. In the combined steering experiment, pairwise modeling raises rank fidelity from 0.5874 to 0.7790 and reduces ordering error from 0.1444 to 0.0606. Together, these results show that individual feature effects need not suffice for faithful composition.

1 INTRODUCTION

Sparse autoencoders (SAEs) decompose model activations into sparse, linearly decoded features. These coordinates are used to inspect computation, build multi-feature explanations, and manipulate behavior. This raises a basic question. If the effects of individual features are known, can their joint effect be predicted by summing them? A linear SAE decoder does not answer this because subsequent computation can make feature effects reinforce, suppress, or depend on one another.

This matters when SAE features are combined into circuits, joint explanations, or interventions (Marks et al., 2025; Ameisen et al., 2025; Ayonrinde et al., 2024). Existing SAE evaluations characterize reconstruction, sparsity, and interpretability (Karvonen et al., 2025; Engels et al., 2025b; Chanin et al., 2025; Engels et al., 2025a), but do not tell us when independently characterized features can be composed into faithful joint explanations or interventions.

To address this gap, we introduce a framework for measuring behavioral compositionality in SAE coordinates. For each prompt, we retain different subsets of selected decoded contributions while holding the remaining activation fixed, producing a downstream response over feature subsets. If individual feature effects compose faithfully, this response should be well described at first order. Departures from that approximation reveal interaction, and adding successively higher interaction orders shows how that missing structure is organized. This lets us compare learned SAE coordinates with matched rotations of the same selected activation subspace, testing whether compositionality under these interventions depends on coordinate choice beyond the subspace and downstream continuation and whether the remaining interaction matters for explanation and steering.

We quantify this structure with SAE Additivity Residuals (SAR). At first order, SAR_1 is the behavioral variation not explained by individual feature effects. Adding pairs gives SAR_2, so $\Delta_2 = \mathrm{SAR}_1 - \mathrm{SAR}_2$ is the structure recovered by pairs and $\rho_2 = \Delta_2/\mathrm{SAR}_1$ its share of the first order residual. Under independent subset sampling, SAR has an exact interaction decomposition so that residual structure above a given order is exactly the normalized energy of higher order interactions. The same quantity also controls errors in preserving the ordering of feature coalitions. Figure 1 illustrates the intervention and contrasts nearly additive and interacting responses.

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.

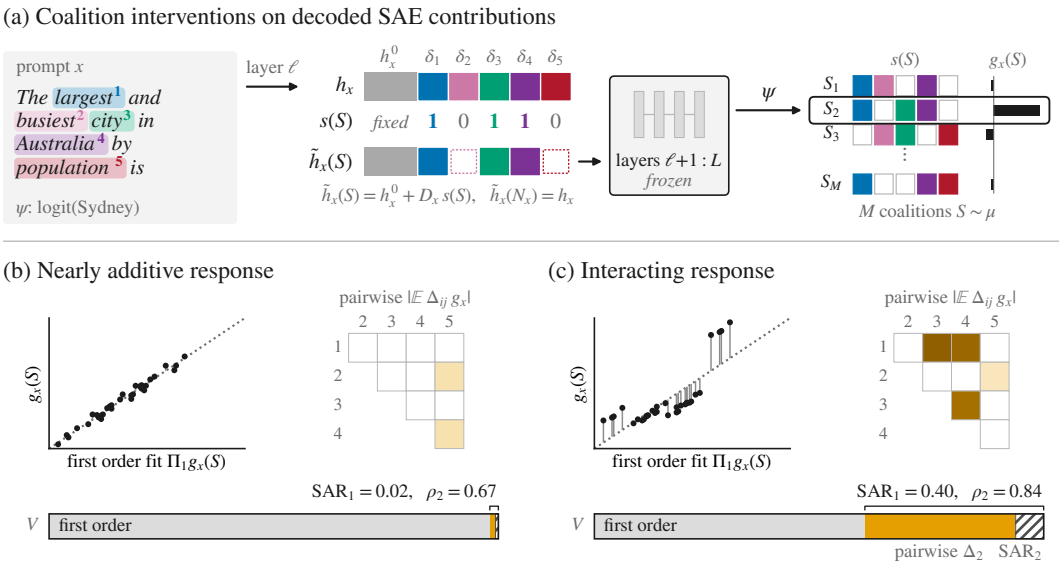


Figure 1: Behavioral compositionality in SAE coordinates. **(a)** Selected decoded contributions are retained or removed while all other activation content is fixed, and the patched activation is propagated through the frozen upper layers to obtain $g_x(S)$. **(b, c)** Nearly additive and interacting responses, showing the first order fit, pairwise interactions, and variance decomposition. SAR_1 is the residual beyond individual features and $\rho_2 = \Delta_2/\mathrm{SAR}_1$ its pairwise share.

The main experiments span three model and SAE pipelines and six behavioral readouts, with additional analyses across seven pipelines and multiple representation depths. We evaluate matched coordinate rotations, first order and pairwise coalition models, combined SAE steering directions, and reconstruction quality. Our contributions are:

- We formulate behavioral compositionality for SAE representations and connect low order approximation error to explicit feature interactions and coalition ordering.
- Using sum preserving rotations of the same selected activation subspace, we show that compositionality under the subset intervention depends on coordinate choice, with learned SAE coordinates supporting a lower order behavioral description than the tested rotations.
- We show that pairwise terms improve held out coalition ordering in each of the 18 tested settings and outperform parameter matched random third order controls.
- We extend the analysis to a combined SAE steering experiment, where pairwise modeling raises rank fidelity from 0.5874 to 0.7790, and find that SAR and reconstruction quality differ in a matched six SAE comparison.

2 Related Work

SAE evaluation commonly measures reconstruction, sparsity, interpretability, and downstream utility (Karvonen et al., 2025). Reconstruction residuals retain structure, concepts can be absorbed across latents, and some model features require multidimensional descriptions (Engels et al., 2025b; Chanin et al., 2025; Engels et al., 2025a). Polynomial SAEs add explicit higher order decoder terms (Koromilas et al., 2026). These works characterize SAE representations and reconstruction, but not whether fixed SAE coordinates compose behaviorally after their decoded contributions pass through the remaining model.

Work on multi-feature explanations includes independent additivity (Ayonrinde et al., 2024), attribution graphs and feature circuits (Ameisen et al., 2025; Marks et al., 2025), and interventions on internal components (Wang et al., 2023; Conmy et al., 2023; Hanna et al., 2024; Kramár et al., 2024). Because circuits, joint explanations, and interventions combine multiple components, individual

characterization does not establish that their effects admit a faithful low order joint description. This motivates measuring compositionality directly in the downstream response.

Interaction indices study joint effects among multiple players (Lundberg & Lee, 2017; Sundararajan et al., 2020; Tsai et al., 2023; Bordt & von Luxburg, 2023). SAEs provide a structured setting in which the players are learned sparse features intended to separate activation structure into interpretable coordinates. We study the interaction that remains when these coordinates are combined downstream and how it is distributed across orders. Appendix C gives the relation to interaction indices and coalition weighting.

3 Behavioral Compositionality of SAE Representations

We first define a controlled response over subsets of selected SAE contributions. We then measure the structure unexplained at a given interaction order, connect it to explicit interactions and coalition ordering, and estimate the resulting population quantities from sampled coalitions.

3.1 From SAE features to controlled subset responses

To study how SAE features combine, we vary selected decoded contributions while holding the rest of the activation fixed. Let F be a language model and $h_x = h_\ell(x)$ the activation at a chosen layer and token position for input x. For an SAE trained at this site, let $z_i(x)$ be feature i's activation and d_i its decoder direction. Let N_x contain the top K active SAE features for prompt x, with decoded contribution

$$\delta_i(x) = z_i(x)d_i, \qquad i \in N_x. \tag{1}$$

For $S \subseteq N_x$, we retain the contributions in S and remove the other selected contributions,

$$\widetilde{h}_x(S) = h_x - \sum_{i \in N_x \setminus S} \delta_i(x) = h_x^0 + \sum_{i \in S} \delta_i(x), \tag{2}$$

where $h_x^0 = h_x - \sum_{i \in N_x} \delta_i(x)$ contains all fixed activation content, including the reconstruction residual, so $\widetilde{h}_x(N_x) = h_x$. The full coalition therefore recovers the original activation, while the reconstruction residual and all unselected content remain fixed, isolating behavioral changes due to the retained selected contributions.

Let D_x contain the selected contributions as columns and $s(S) \in \{0,1\}^K$ indicate those retained, so

$$\widetilde{h}_x(S) = h_x^0 + D_x s(S). \tag{3}$$

For a scalar behavior ψ, the downstream response is

$$g_x(S) = \psi\Big(F_{\ell \leftarrow \widetilde{h}_x(S)}(x)\Big), \qquad S \subseteq N_x. \tag{4}$$

This defines a controlled set-function response over selected SAE coordinates while the activation remainder and downstream model remain fixed. The object of analysis is downstream behavior rather than SAE reconstruction.

To compare learned coordinates with alternatives spanning the same selected subspace, let $Q \in \mathbb{R}^{K \times K}$ be orthogonal with $Q\mathbf{1} = \mathbf{1}$ and define

$$D_x^Q = D_x Q. \tag{5}$$

Orthogonality preserves the selected subspace and $Q\mathbf{1} = \mathbf{1}$ its total contribution; the fixed remainder and downstream model are unchanged. This gives a sum preserving random coordinate baseline within the same selected subspace.

3.2 Measuring interaction beyond individual feature effects

We first ask whether individual feature effects suffice to describe g_x, and if not, what interaction order is required. For fixed x, write $N = N_x$ and $|N| = K$. We approximate g_x using effects involving at

most r selected features, with function class

$$\mathcal{A}_r(N) = \left\{ a(S) = \sum_{\substack{B \subseteq N \\ |B| \le r}} \theta_B \mathbf{1}\{B \subseteq S\} \right\}. \tag{6}$$

Thus $r = 1$ contains individual feature effects, while $r = 2$ additionally includes feature pairs.

Let μ denote the distribution over feature subsets and $\Pi_r g_x$ the $L^2(\mu)$ projection of g_x onto $\mathcal{A}_r(N)$, which minimizes $\|g_x - a\|_\mu^2$. Define

$$V(x;\psi) = \|g_x - \mathbb{E}_{S\sim\mu} g_x(S)\|_\mu^2 . \tag{7}$$

The SAE Additivity Residual is

$$\mathrm{SAR}_r(x;\psi) = \begin{cases} \dfrac{\|g_x - \Pi_r g_x\|_\mu^2}{V(x;\psi)}, & V(x;\psi) > 0, \\ 0, & V(x;\psi) = 0. \end{cases} \tag{8}$$

Thus SAR_r is the fraction of behavioral variance unexplained by the best order r approximation; SAR_1 is the residual beyond individual feature effects.

SAR is an approximation residual, but under independent subset sampling an orthogonal product representation identifies this residual exactly with higher order interaction structure. Under independent feature inclusion with probability $p \in (0,1)$, define

$$u_i(S) = \frac{\mathbf{1}\{i \in S\} - p}{\sqrt{p(1-p)}}, \qquad u_B(S) = \prod_{i \in B} u_i(S), \tag{9}$$

$c_B = \mathbb{E}[g_x(S) u_B(S)]$, and for $S' \subseteq N \setminus B$,

$$\Delta_B g_x(S') = \sum_{C \subseteq B} (-1)^{|B|-|C|} g_x(S' \cup C). \tag{10}$$

For $B = \{i, j\}$, $\Delta_{ij} g_x(S')$ is the pairwise effect on background S'; $c_{\{i,j\}}$ averages its signed value over backgrounds.

Theorem 1 (Behavioral interaction decomposition)**.** *If coalition memberships are independent Bernoulli variables with probability $p \in (0,1)$, then*

$$c_B = [p(1-p)]^{|B|/2} \mathbb{E}[\Delta_B g_x(S')], \tag{11}$$

and, when $V(x;\psi) > 0$,

$$\mathrm{SAR}_r(x;\psi) = \frac{\sum_{|B|>r} [p(1-p)]^{|B|} \left(\mathbb{E}[\Delta_B g_x]\right)^2}{V(x;\psi)}. \tag{12}$$

Under independent subset sampling, SAR_r is exactly the fraction of behavioral variance in product measure interaction terms above order r; SAR_1 aggregates all such pairwise and higher order terms.

The main analysis uses the uniform distribution over all subsets, equivalent to $p = 1/2$. Then SAR_1 measures interaction beyond individual coordinates and

$$\begin{aligned} \Delta_2(x;\psi) &= \mathrm{SAR}_1(x;\psi) - \mathrm{SAR}_2(x;\psi) \\ &= \frac{1}{16 V(x;\psi)} \sum_{i<j} \left(\mathbb{E}[\Delta_{ij} g_x]\right)^2 , \end{aligned} \tag{13}$$

with pairwise share

$$\rho_2(x;\psi) = \frac{\Delta_2(x;\psi)}{\mathrm{SAR}_1(x;\psi)}. \tag{14}$$

for $\mathrm{SAR}_1 > 0$, and set $\rho_2 = 0$ when $\mathrm{SAR}_1 = 0$. Here SAR_1 measures interaction beyond individual effects, Δ_2 the part recovered by pairs, and SAR_2 what remains beyond pairwise order. Thus ρ_2 is the pairwise share in this product measure decomposition.

The projection definition also gives several basic calibration properties:

$$\mathrm{SAR}_r = 0 \iff g_x \in \mathcal{A}_r(N_x), \qquad 0 = \mathrm{SAR}_K \le \cdots \le \mathrm{SAR}_1 \le 1, \tag{15}$$

and, for $a \in \mathcal{A}_r(N_x)$,

$$\frac{\|g_x - a\|_\mu^2}{V(x;\psi)} \ge \mathrm{SAR}_r(x;\psi). \tag{16}$$

Thus SAR decreases with interaction order, vanishes exactly when order r is sufficient, and leaves explained variance $1 - \mathrm{SAR}_r$. Proofs are in Appendix A.

First order additivity is guaranteed when the downstream continuation is affine over the tested activation states.

Corollary 1 (Affine continuation). *If $b_x(h) = \psi(F_{\ell \leftarrow h}(x)) = c + w^\top h$ for every tested $h = \widetilde{h}_x(S)$, then $\mathrm{SAR}_1(x;\psi) = 0$.*

Thus nonzero SAR_1 reflects departure from affine continuation on the tested SAE states.

3.3 Connecting low order approximation to coalition ordering

Joint explanations should also preserve comparisons between feature coalitions. Let S, T be independent coalitions and write $\Delta g = g_x(S) - g_x(T)$ and $\Delta a = a(S) - a(T)$. Define

$$\mathcal{E}(a) = \frac{\mathbb{E}[(\Delta g)^2 \mathbf{1}\{\Delta g\, \Delta a \le 0\}]}{2V(x;\psi)}. \tag{17}$$

The weighting emphasizes reversals between subsets with larger behavioral differences and down-weights near ties.

Theorem 2 (Coalition ordering error). *For any function a on 2^N and $V(x;\psi) > 0$,*

$$\mathcal{E}(a) \le \frac{\mathrm{Var}_\mu(g_x - a)}{V(x;\psi)} \le \frac{\|g_x - a\|_\mu^2}{V(x;\psi)}. \tag{18}$$

In particular,

$$\mathcal{E}(\Pi_r g_x) \le \mathrm{SAR}_r(x;\psi). \tag{19}$$

Thus small SAR_r guarantees small behavior weighted ordering error for the optimal order r approximation, although the converse need not hold. We examine below its association with held out coalition ranking.

3.4 Estimating compositionality from sampled coalitions

Because exhaustive evaluation scales as 2^K, we fit low order models on sampled coalitions and evaluate them on held out subsets. We sample from μ, split into fitting and evaluation sets, and fit

$$\widehat{a}_r = \operatorname*{arg\,min}_{a \in \mathcal{A}_r(N_x)} \frac{1}{M_{\mathrm{fit}}} \sum_{m=1}^{M_{\mathrm{fit}}} \left(g_x(S_m) - a(S_m)\right)^2. \tag{20}$$

On the evaluation set, let $\widehat{L}_r$ be the mean squared error and $\widehat{V}$ the empirical variance of g_x. We report

$$\widehat{\mathrm{SAR}}_r = \frac{\widehat{L}_r}{\widehat{V} + \epsilon}. \tag{21}$$

The fitted residual and population quantity are related by

$$\|g_x - \widehat{a}_r\|_\mu^2 = \|g_x - \Pi_r g_x\|_\mu^2 + \|\widehat{a}_r - \Pi_r g_x\|_\mu^2. \tag{22}$$

The two terms are structural residual and finite sample fitting error, so $\widehat{\mathrm{SAR}}_r$ is a held out predictive residual rather than population SAR itself.

For the evaluation set $\mathcal{U}$, we also report Spearman rank fidelity,

$$F_{\mathrm{rank}}^{(r)} = \mathrm{Spearman}_{S \in \mathcal{U}}\left(\widehat{a}_r(S), g_x(S)\right), \tag{23}$$

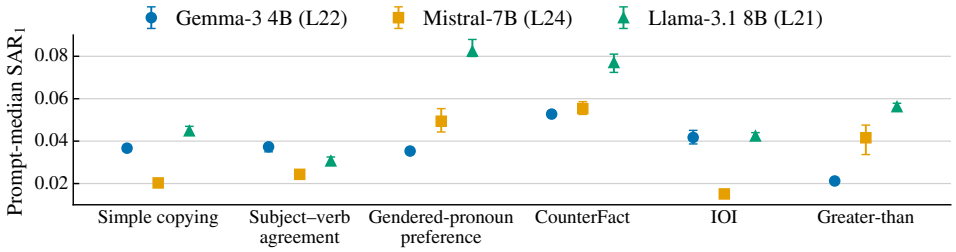


Figure 2: Prompt median $\widehat{\mathrm{SAR}}_1$ for six behaviors in Gemma-3 4B layer 22, Mistral-7B layer 24, and Llama-3.1 8B layer 21. Error bars show 95% prompt bootstrap intervals.

and denote the held out analogue of Equation (17) by $\widehat{\mathcal{E}}_r$. We additionally report sign error and selection regret. With $e(S) = g_x(S) - g_x(N_x)$ and $\widehat{e}_r(S) = \widehat{a}_r(S) - \widehat{a}_r(N_x)$, sign error is the fraction of held out coalitions with $e(S)\widehat{e}_r(S) < 0$. If $S^\star = \arg\max_{S\in\mathcal{U}} g_x(S)$ and $\widehat{S}_r = \arg\max_{S\in\mathcal{U}} \widehat{a}_r(S)$, selection regret is

$$R^{(r)} = \frac{g_x(S^\star) - g_x(\widehat{S}_r)}{\max_{S\in\mathcal{U}} g_x(S) - \min_{S\in\mathcal{U}} g_x(S) + \epsilon}.$$

All predictive metrics use coalitions unseen during fitting. Appendix B reports conditioning, split and sampling stability, and fitting budget diagnostics.

4 Experiments

We first characterize compositionality across behaviors and representations, then compare learned coordinates with matched random rotations and test whether pairwise structure recovers failures of first order explanations. We finally examine depth, combined steering directions, and the relation to reconstruction quality. The main comparisons use Gemma-3 4B, Mistral-7B, and Llama-3.1 8B across simple copying, subject verb agreement, gendered pronoun preference, CounterFact (Meng et al., 2022), IOI (Wang et al., 2023), and greater than comparison. The fixed layer landscape, rotation, fidelity, and pairwise analyses use 1,000, 100, 800, and 150 prompts per model and behavior setting, respectively. For each prompt, we select the top $K = 16$ active SAE contributions and sample 384 coalitions, using 256 for fitting and 128 for evaluation. Appendix D gives the full model, behavior, and sampling specifications.

4.1 Behavioral compositionality varies across behaviors and representations

We first ask whether first order additivity is a reasonable approximation across behaviors and representations. The prompt median $\widehat{\mathrm{SAR}}_1$ is positive in all 57 model, layer, and behavior combinations while affine controls remain at numerical zero. Its magnitude also varies across settings. Simple copying and subject verb agreement have lower residuals than CounterFact in all three main pipelines, while the other behaviors change order across models. Thus first order additivity is not exact in these intervention games, and the estimated residual varies with both behavior and representation.

4.2 Compositionality depends on coordinate choice under subset interventions

We compare each SAE basis with sum preserving rotations of the same selected subspace to test whether compositionality under the subset intervention depends on coordinate choice rather than only on the subspace and downstream continuation. Figure 3 shows the paired comparison for all 18 model and behavior combinations. Learned SAE coordinates have lower median $\widehat{\mathrm{SAR}}_1$ in every comparison, although the size of the gap varies across settings. For example, the medians are 0.0218 and 0.1196 for Mistral subject verb agreement. The selected subspace, total contribution, and downstream continuation remain fixed as the intervention coordinates change. The systematic SAR gap therefore shows that, under this intervention, these shared factors alone do not determine compositionality; the learned SAE basis supports a lower order description than the tested rotations. Under milder interventions,

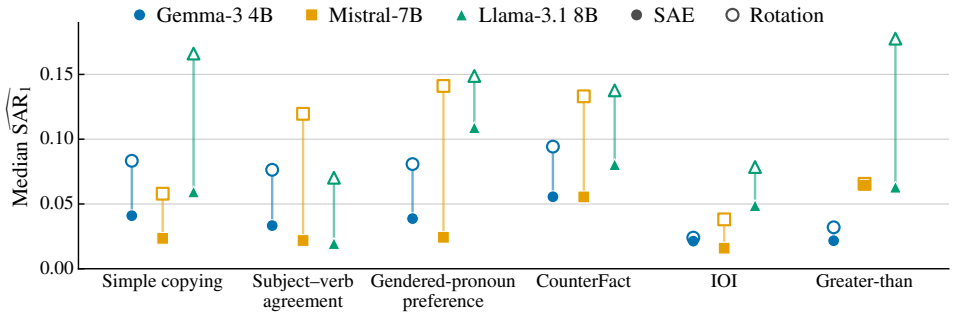


Figure 3: Paired first order SAR for learned SAE coordinates and sum preserving random rotations of the same selected activation subspace.

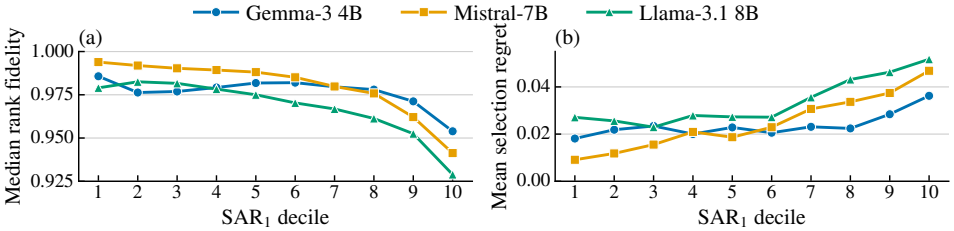


Figure 4: First order explanation fidelity across $\widehat{\mathrm{SAR}}_1$ deciles within each model.

pairwise rank improvement persists at $p = 0.75$, while halving removed contributions lowers $\widehat{\mathrm{SAR}}_1$ and improves first order rank fidelity in eight of nine settings. Appendix F gives the full controls.

4.3 Estimated SAR is associated with first order coalition fidelity

The ordering theorem bounds the population behavior weighted ordering error. As a descriptive calibration, we compare estimated SAR with rank fidelity and selection regret on the same held out coalitions. Figure 4 shows lower rank fidelity and higher selection regret from the lowest to the highest $\widehat{\mathrm{SAR}}_1$ decile in all three models. Selection regret approximately doubles in Gemma and Llama and increases by about fivefold in Mistral. Across 18 model and behavior combinations, the prompt level correlation between $\widehat{\mathrm{SAR}}_1$ and rank fidelity is negative in 17 and at most -0.6 in 14.

4.4 Feature pairs recover missing coalition structure

The previous results identify settings where individual feature effects are insufficient. We therefore ask how much of the missing structure can be recovered by the next interaction order, namely feature pairs. Figure 5 shows the paired comparison between first order and pairwise approximations across all 18 model and behavior combinations. Pairwise models improve rank fidelity in all 18 model and behavior combinations, raising the average from 0.9782 to 0.9959, and satisfy $\widehat{\mathcal{E}}_2 < \widehat{\mathcal{E}}_1$ in all 18. Both improvements persist across all 20 repeated fitting and evaluation splits. Selection regret falls from 0.0272 to 0.0051 on average, and sign error also decreases in every combination. These results show that pairwise terms capture substantial predictive structure beyond individual feature effects. Full results are in Appendix H.

A parameter matched control retains the first order terms and replaces the pairwise terms with the same number of randomly selected third order terms. Pairwise models have higher held out rank fidelity and lower behavior weighted ordering error than the median matched model in all 18 combinations. The full comparison is in Appendix H.4.

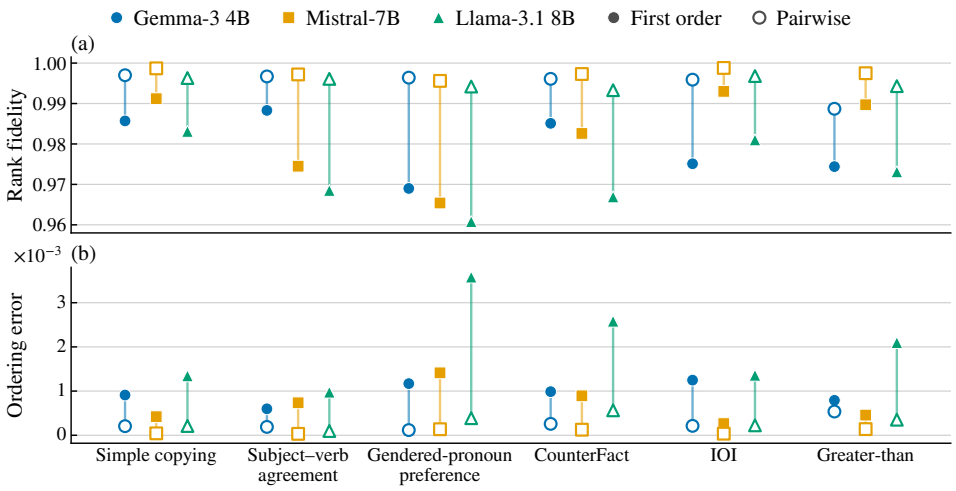


Figure 5: Held out coalition ordering for first order and pairwise approximations.

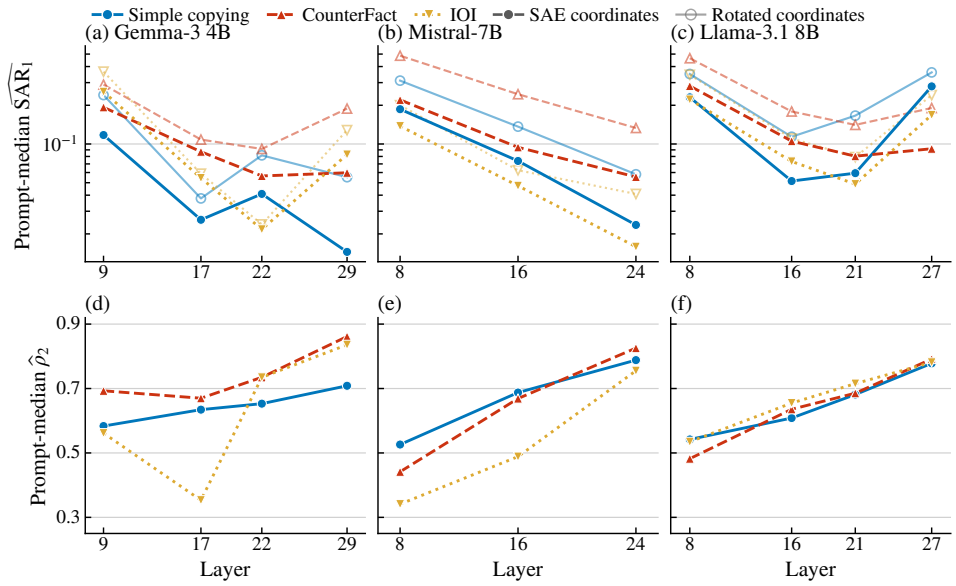


Figure 6: Depth profiles for simple copying, CounterFact, and IOI.

4.5 Estimated compositionality varies across depth

We next examine how the estimated compositionality measures change across representation depth. Figure 6 follows simple copying, CounterFact, and IOI. Across nine early to late comparisons, $\widehat{\mathrm{SAR}}_1$ decreases in eight and $\widehat{\rho}_2$ increases in all nine. Later representations therefore usually have lower estimated first order residuals, while $\widehat{\rho}_2$ increases. This may reflect either the SAE coordinates or the shorter remaining downstream computation. Random rotations show the same first order decrease in eight of nine comparisons, indicating that part of the depth trend comes from the remaining downstream computation rather than the SAE basis. SAE coordinates nevertheless have lower $\widehat{\mathrm{SAR}}_1$ than rotations in all 33 model, behavior, and layer comparisons. Thus part of the depth trend is shared across coordinate systems, while SAE coordinates remain lower residual than rotations at matched sites; $\widehat{\rho}_2$ also increases. The Llama profiles are nonmonotonic near the final sampled layer.

Across all seven pipelines, simple copying has lower $\widehat{\mathrm{SAR}}_1$ at the latest sampled representation in six cases and higher $\widehat{\rho}_2$ in all seven. The pattern therefore extends beyond the three main pipelines. Appendix Figure 7 and Table 13 give the full comparison.

4.6 Pairwise modeling better predicts the aggregate steering response

The preceding experiments study interactions among feature contributions already present in a model activation. SAE features are also used by adding directions to steer model behavior. If these effects compose poorly, effects measured for individual steering directions may not predict the outcome of combining them. We therefore apply the same analysis to fixed published steering directions. Using 16 published CorrSteer features (Cho et al., 2026) in Llama-3.1 8B, we evaluate 1,024 steering combinations with separate fitting and evaluation coalitions. On the overall 1,400 question MMLU-Pro response (Wang et al., 2024), rank fidelity rises from 0.5874 to 0.7790, behavior weighted ordering error falls from 0.1444 to 0.0606, and the held out residual estimate decreases from 0.6424 to 0.3891. Pairwise modeling therefore better predicts this aggregate steering response; domain level effects are heterogeneous and reported in Appendix H.

Table 1: Held out residual estimates and coalition ordering for MMLU-Pro steering response.

Approximation	Residual estimate	Rank fidelity	Ordering error
First order	0.6424	0.5874	0.1444
Pairwise	0.3891	0.7790	0.0606

4.7 Reconstruction quality and SAR differ in a matched SAE comparison

We next ask whether reconstruction quality tracks behavioral compositionality across matched SAEs at a fixed site. Pooling prompts across six SAEs at a fixed Gemma-3 1B site, SAR is positively related to sign error and selection regret in all three behaviors, while FVU_x is negative or near zero for sign error and near zero for selection regret. Thus reconstruction quality and SAR capture different variation in joint behavioral fidelity in this matched comparison. Appendix Figure 8 gives the SAE comparison; Appendix I reports sensitivity to K.

Table 2: Prompt level Spearman correlations pooled over six SAEs ($n = 1{,}800$ per behavior).

	Sign error		Selection regret	
Behavior	FVU_x	$\widehat{\mathrm{SAR}}_1$	FVU_x	$\widehat{\mathrm{SAR}}_1$
Simple copying	-0.080	0.269	0.011	0.220
Subject verb agreement	-0.114	0.430	0.005	0.235
IOI	-0.013	0.396	0.028	0.199

5 Discussion and Conclusion

Linear decoding makes SAE coordinates additive for activation reconstruction, not for downstream effects. Reconstruction quality, individual feature interpretability, and the interaction order needed for joint behavior are therefore distinct representation properties, with SAR measuring the last directly. This matters whenever SAE features are combined: meaningful individual coordinates need not have additive joint behavioral effects. Learned SAE coordinates nevertheless have lower first order residuals than the tested sum preserving random rotations of the same selected contribution subspaces. Individual effects remain insufficient in many cases, while pairwise terms improve coalition prediction. The SAE basis therefore supports a lower order description than these rotations without making feature effects independent, so residual interaction can matter for faithful joint explanations. The distinction also appears in steering and differs from reconstruction quality in the matched SAE comparison. These conclusions depend on the selected contributions, behavioral readout, and intervention distribution rather than a global property of an SAE dictionary. Behavioral compositionality is therefore a distinct property to consider when SAE features are combined for explanation or intervention.

AI USE STATEMENT

In this work, we used generative AI tools to improve the readability of the manuscript, to check the correctness and internal consistency of mathematical derivations and proofs, and to assist with experimental plotting.

REPRODUCIBILITY STATEMENT

Population SAR, its held out estimator, the interaction decomposition, the coalition ordering result, and the projection estimator are defined in Section 3. Their proofs and the finite sample least squares analysis appear in Appendices A and B. The experimental protocol, including behavior definitions, model and SAE coverage, feature set size, coalition sampling, and fitting and evaluation splits, is specified in Section 4 and Appendix D. The main experiments use $K = 16$, 384 sampled coalitions per prompt, and separate fitting and evaluation sets. The steering experiment uses 1,024 sampled coalitions with a fixed equal split, and the coordinate and parameter matched comparisons use paired coalition assignments. Measurement calibration appears in Appendix E, and sensitivity and protocol checks appear in Appendix I.

REFERENCES

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, et al. Circuit tracing: Revealing computational graphs in language models. Transformer Circuits Thread, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Kola Ayonrinde, Michael T. Pearce, and Lee Sharkey. Interpretability as compression: Reconsidering SAE explanations of neural activations with MDL-SAEs. *arXiv preprint arXiv:2410.11179*, 2024.

Joseph Bloom, Curt Tigges, Anthony Duong, and David Chanin. SAELens. https://github.com/decoderesearch/SAELens, 2024.

Sebastian Bordt and Ulrike von Luxburg. From shapley values to generalized additive models and back. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206 of *Proceedings of Machine Learning Research*, pp. 709–745, 2023. URL https://arxiv.org/abs/2209.04012.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *Advances in Neural Information Processing Systems*, volume 38, pp. 82318–82355, 2025. URL https://arxiv.org/abs/2409.14507.

Seonglae Cho, Zekun Wu, and Adriano Koshiyama. CorrSteer: Generation-time LLM steering via correlated sparse autoencoder features. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2508.12535.

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*, volume 36, pp. 16318–16352, 2023.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025a. URL https://arxiv.org/abs/2405.14860.

Joshua Engels, Logan Riggs Smith, and Max Tegmark. Decomposing the dark matter of sparse autoencoders. *Transactions on Machine Learning Research*, 2025b. URL https://arxiv.org/abs/2410.14670.

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28(4):547–565, 1999.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *First Conference on Language Modeling*, 2024.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. Llama scope: Extracting millions of features from Llama-3.1-8B with sparse autoencoders. *arXiv preprint arXiv:2410.20526*, 2024. URL https://arxiv.org/abs/2410.20526.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29223–29264, 2025. URL https://proceedings.mlr.press/v267/karvonen25a.html.

Panagiotis Koromilas, Andreas D. Demou, James Oldfield, Yannis Panagakis, and Mihalis A. Nicolaou. Polysae: Modeling feature interactions in sparse autoencoders via polynomial decoding. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://arxiv.org/abs/2602.01322.

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.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4765–4774, 2017.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Callum McDougall, Arthur Conmy, János Kramár, Tom Lieberum, Senthooran Rajamanoharan, and Neel Nanda. Gemma scope 2: Technical paper. Technical report, Google DeepMind, 2025. URL https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/gemma-scope-2-helping-the-ai-safety-community-deepen-understanding-of-complex-language-model-behavior/Gemma_Scope_2_Technical_Paper.pdf.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games, Volume II*, pp. 307–317. Princeton University Press, 1953.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The Shapley Taylor interaction index. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9259–9268, 2020.

Che-Ping Tsai, Chih-Kuan Yeh, and Pradeep Ravikumar. Faith-Shap: The faithful Shapley interaction index. *Journal of Machine Learning Research*, 24(94):1–42, 2023.

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 *The Eleventh International Conference on Learning Representations*, 2023.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290, 2024. doi: 10.52202/079017-3018. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract.html.

A Proofs and Mathematical Details

A.1 Notation and equivalent bases

Table 3: Notation used throughout the paper.

Symbol	Meaning
x	Input prompt.
h_x	Model activation at the selected layer and token position.
$z_i(x), d_i$	Activation and decoder direction of SAE feature i.
$\delta_i(x)$	Decoded contribution $z_i(x)d_i$.
N_x	Set of top K active SAE features for prompt x.
S	Coalition of retained features, $S \subseteq N_x$.
g_x	SAE feature intervention game.
μ	Coalition distribution.
$\mathcal{A}_r(N_x)$	Set functions with interaction order at most r.
$V(x;\psi)$	Behavioral variance across interventions.
$\mathrm{SAR}_r(x;\psi)$	Population residual after the order r projection.
$\widehat{\mathrm{SAR}}_r$	Held out residual estimate from the fitted order r model.
$\Delta_B g_x$	Finite difference associated with feature subset B.
Δ_2, ρ_2	Population pairwise component and its share of SAR_1.
$\widehat{\Delta}_2, \widehat{\rho}_2$	Corresponding held out estimates.
$\mathcal{E}(a)$	Population behavior weighted coalition ordering error.
$\widehat{\mathcal{E}}_r$	Ordering error on held out coalitions.

The indicator functions $\mathbf{1}\{B \subseteq S\}$ with $|B| \le r$ span $\mathcal{A}_r(N)$. Under product Bernoulli sampling, the functions

$$u_i(S) = \frac{\mathbf{1}\{i \in S\} - p}{\sqrt{p(1-p)}}, \qquad u_B(S) = \prod_{i \in B} u_i(S) \tag{24}$$

form an orthonormal basis of $L^2(2^N, \mu_p)$. The span of $\{u_B \mid |B| \le r\}$ is the same space $\mathcal{A}_r(N)$. At $p = 1/2$, the two coordinate systems are related by the invertible transformation $\mathbf{1}\{i \in S\} = (1 + u_i(S))/2$. This equivalence is used in the finite sample analysis, while the finite sample estimator is expressed in the indicator basis.

A.2 Proof of the behavioral interaction decomposition

Fix $B \subseteq N$ and condition on the memberships of the features outside B. Let the resulting coalition be $S' \subseteq N \setminus B$. For each $i \in B$, a present feature contributes

$$p\frac{1-p}{\sqrt{p(1-p)}} = \sqrt{p(1-p)} \tag{25}$$

to the conditional expectation of u_i, while an absent feature contributes

$$(1-p)\frac{-p}{\sqrt{p(1-p)}} = -\sqrt{p(1-p)}. \tag{26}$$

Expanding over the memberships inside B therefore gives

$$\mathbb{E}[g_x(S)u_B(S) \mid S'] \tag{27}$$

$$= [p(1-p)]^{|B|/2} \sum_{C \subseteq B} (-1)^{|B|-|C|} g_x(S' \cup C) \tag{28}$$

$$= [p(1-p)]^{|B|/2} \Delta_B g_x(S'). \tag{29}$$

Averaging over S' proves Equation (11).

Because the u_B form an orthonormal basis,

$$g_x(S) = \sum_{B \subseteq N} c_B u_B(S), \qquad \|g_x - \Pi_r g_x\|^2_{\mu_p} = \sum_{|B| > r} c_B^2. \tag{30}$$

The centered variance is

$$V(x;\psi) = \sum_{|B|\geq 1} c_B^2. \tag{31}$$

Substituting the finite difference expression for c_B proves Theorem 1 and Equation (13).

A.3 Projection properties

If $g_x \in \mathcal{A}_r$, then $\Pi_r g_x = g_x$ and $\mathrm{SAR}_r = 0$. Conversely, if $\mathrm{SAR}_r = 0$, full support makes g_x equal to its projection on every coalition. The nesting

$$\mathcal{A}_0 \subseteq \mathcal{A}_1 \subseteq \cdots \subseteq \mathcal{A}_K \tag{32}$$

then gives the residual sequence in Equation (15).

For any $a \in \mathcal{A}_r$, projection orthogonality gives

$$\|g_x - a\|_\mu^2 = \|g_x - \Pi_r g_x\|_\mu^2 + \|\Pi_r g_x - a\|_\mu^2, \tag{33}$$

which proves Equation (16). Since constants belong to $\mathcal{A}_r$, the projection residual has mean zero. Applying the same orthogonal decomposition after centering gives

$$\frac{\mathrm{Var}_\mu(\Pi_r g_x)}{V(x;\psi)} = 1 - \mathrm{SAR}_r(x;\psi). \tag{34}$$

If the continuation map is affine on the tested states,

$$g_x(S) = c + w^\top \left(h_x^0 + \sum_{i\in S} \delta_i(x) \right) \tag{35}$$

$$= c + w^\top h_x^0 + \sum_{i\in S} w^\top \delta_i(x), \tag{36}$$

so $g_x \in \mathcal{A}_1$ and Corollary 1 follows. SAR is also invariant to affine rescaling of g_x because both its numerator and denominator scale by the same squared factor.

A.4 Proof of the coalition ordering theorem

Let

$$e = g_x - a, \qquad \Delta e = e(S) - e(T) = \Delta g - \Delta a. \tag{37}$$

If $\Delta g\, \Delta a \leq 0$ and $\Delta g \neq 0$, then $|\Delta e| \geq |\Delta g|$. Hence

$$(\Delta g)^2 \mathbf{1}\{\Delta g\, \Delta a \leq 0\} \leq (\Delta e)^2. \tag{38}$$

For independent $S, T \sim \mu$,

$$\mathbb{E}[(\Delta e)^2] = 2\,\mathrm{Var}_\mu(e), \qquad \mathbb{E}[(\Delta g)^2] = 2V(x;\psi). \tag{39}$$

Dividing the first inequality by the second expectation gives the first bound in Equation (18). The second follows from $\mathrm{Var}(e) \leq \mathbb{E}[e^2]$. When $a = \Pi_r g_x$, the residual has mean zero and its normalized squared norm is SAR_r, proving Equation (19).

B Estimation and Finite Sample Analysis

B.1 Projection estimator

For each prompt, we draw coalitions with independent feature memberships and evaluate $g_x(S_m)$. For order r, we use

$$\phi_r(S) = (\mathbf{1}\{B \subseteq S\})_{B\subseteq N_x,\, |B|\leq r} \tag{40}$$

and fit $\widehat{a}_r(S) = \widehat{\theta}_r^\top \phi_r(S)$ by unregularized least squares on the fitting coalitions. On the evaluation split we compute

$$\widehat{L}_r = \frac{1}{M_{\mathrm{eval}}} \sum_{m=1}^{M_{\mathrm{eval}}} \left(g_x(S_m) - \widehat{a}_r(S_m)\right)^2 \tag{41}$$

and the held out variance $\widehat{V}$. The reported estimator is

$$\widehat{\mathrm{SAR}}_r = \frac{\widehat{L}_r}{\widehat{V} + \epsilon}. \tag{42}$$

The three large model pipelines use $\epsilon = 10^{-8}$ and float64 projection.

B.2 Walsh coordinates and fitting error

At $p = 1/2$, let $\varphi_r(S)$ contain the Walsh basis terms $u_B(S)$ with $|B| \le r$. Write

$$a_r^*(S) = \Pi_r g_x(S) = \theta_*^\top \varphi_r(S), \qquad \varepsilon_r(S) = g_x(S) - a_r^*(S). \tag{43}$$

The population Gram matrix is the identity and $\mathbb{E}[\varphi_r(S)\varepsilon_r(S)] = 0$. For a fitting sample of size M, define

$$\widehat{G}_r = \frac{1}{M} \sum_{m=1}^{M} \varphi_r(S_m)\varphi_r(S_m)^\top. \tag{44}$$

When $\widehat{G}_r$ is invertible, the least squares normal equations give the exact identity

$$\widehat{\theta}_r - \theta_* = \widehat{G}_r^{-1} \frac{1}{M} \sum_{m=1}^{M} \varphi_r(S_m)\varepsilon_r(S_m). \tag{45}$$

Because the population Walsh basis is orthonormal,

$$\|\widehat{a}_r - a_r^*\|_\mu^2 = \|\widehat{\theta}_r - \theta_*\|_2^2. \tag{46}$$

Equations (45) and (46) show how finite sample fitting error depends on the empirical Gram matrix and the sampled projection residual. Combining these identities with Equation (22) gives a finite sample description of the unregularized least squares estimator used in the experiments.

The monomial and Walsh coordinates span the same function class, so a full rank fit gives the same predictions in either basis. The Walsh form is used here only to expose the population orthogonality and conditioning of the sampled problem.

B.3 Finite sample diagnostics

We evaluated the fitting protocol on the 2,700 prompts used in the explanation fidelity experiments. The first and second order models use 17 and 137 coefficients. Every sampled design is full rank. The median condition number of the second order Walsh Gram matrix is 41.02 and its 95th percentile is 49.88. Predictions from the monomial and Walsh coordinates agree to numerical precision.

Table 4: Finite sample diagnostics for the 256 fitting and 128 evaluation coalition protocol. Split variation is the standard deviation across 20 random splits divided by the corresponding split mean. The final two rows compare a 256 coalition fit with a fit using all 384 sampled coalitions and normalize the mean squared prediction difference by behavioral variance.

Diagnostic	Result
Maximum monomial and Walsh prediction difference	1.67×10^{-12}
Order 1 full rank prompts	2,700 / 2,700
Order 2 full rank prompts	2,700 / 2,700
Order 2 Gram condition number, median	41.02
Order 2 Gram condition number, 95th percentile	49.88
Relative split variation of first order residual, median	2.29%
Relative split variation of second order residual, median	2.70%
Cells with pairwise rank improvement in all 20 splits	18 / 18
Cells with pairwise $\mathcal{E}$ reduction in all 20 splits	18 / 18
Median difference from the 384 coalition fit, order 1	6.91×10^{-4}
Median difference from the 384 coalition fit, order 2	1.70×10^{-3}

The fitting budget analysis shows that the pairwise conclusion becomes stable before the full fitting budget even though the precise second order residual continues to change. At 224 and 256 fitting

coalitions, pairwise models improve both rank fidelity and behavior weighted ordering error in all 18 cells. The second order rank fidelity changes by 0.14% from 224 to 256 fitting coalitions, while the normalized second order residual changes by about 21%. We therefore use the repeated split analysis to support the stability of the comparative conclusions and do not interpret the fitted second order residual as a high precision estimate of its population value.

We also generated five independent sets of 384 coalitions for 50 fixed prompts in each of nine representative model and behavior combinations. All nine combinations preserve the original qualitative first order and pairwise conclusions across fresh coalition draws. The median within prompt coefficient of variation is 13.7% for $\widehat{\mathrm{SAR}}_1$ and 18.2% for $\widehat{\mathrm{SAR}}_2$, compared with 0.37% and 0.16% for first order and pairwise rank fidelity. The ordering conclusions are therefore considerably more stable to coalition sampling than the absolute residual magnitudes.

Table 5: Fitting budget diagnostics with the original evaluation coalitions held fixed. The reported values aggregate the 18 model and behavior combinations.

Fit coalitions	Condition	R_2^{var}	$F_{\mathrm{rank}}^{(2)}$	$\mathcal{E}_2$	Rank improved
160	580.3	0.03071	0.9783	0.001085	9 / 18
192	134.3	0.01516	0.9885	0.0003920	17 / 18
224	65.43	0.01104	0.9915	0.0002436	18 / 18
256	41.02	0.009152	0.9929	0.0001839	18 / 18

B.4 Prompt aggregation and paired endpoints

We compute $\widehat{\mathrm{SAR}}_1$, $\widehat{\mathrm{SAR}}_2$, $\widehat{\Delta}_2$, and $\widehat{\rho}_2$ separately for each prompt. Dataset summaries use prompt level medians unless otherwise stated. Rank fidelity, sign error, and selection regret in the primary pairwise comparison use prompt means. Confidence intervals resample prompts with replacement 1,000 times. Endpoint comparisons apply each bootstrap draw to the same prompt IDs at the early and late layers.

C Interaction Indices and Alternative Constructions

C.1 Interaction indices and low order fits

Interaction indices extend feature attribution from individual players to small coalitions. The Shapley value assigns each player a weighted average of its marginal contributions (Shapley, 1953), and its extensions assign values to pairs and larger subsets (Grabisch & Roubens, 1999; Sundararajan et al., 2020; Tsai et al., 2023). Several such quantities can be interpreted as coefficients of low order fits (Lundberg & Lee, 2017; Tsai et al., 2023; Bordt & von Luxburg, 2023). SAR instead aggregates the residual energy that remains above a chosen order.

Under product sampling with $p = 1/2$, the relation to pairwise interactions is explicit. Theorem 1 gives

$$c_{\{i,j\}} = \tfrac{1}{4}\,\mathbb{E}_{S' \subseteq N \setminus \{i,j\}}[\Delta_{ij} g(S')]. \tag{47}$$

Thus the pairwise Fourier coefficient is a scaled context averaged second finite difference. The quantity Δ_2 aggregates the squared pairwise coefficients rather than assigning an individual importance score to each pair.

Low order coefficients alone do not determine the residual above that order. Under the product measure, the basis terms with $|B| \le r$ and $|B| > r$ occupy orthogonal subspaces. Changing the latter leaves every fitted coefficient up to order r unchanged while changing SAR_r. Estimating the residual therefore requires behavioral evaluations beyond the fitted coefficients themselves, which is why the estimator uses a separate evaluation split.

C.2 Coalition weighting and variance shares

The weighting that makes a first order fit reproduce the Shapley value places weight proportional to $[\binom{K}{|S|}|S|(K-|S|)]^{-1}$ on nontrivial coalitions (Lundberg & Lee, 2017). This weighting correlates

feature memberships, so the product basis used in Theorem 1 is no longer orthogonal. The differences in the SAR sequence therefore do not have the same variance share interpretation under that weighting. Product sampling gives independent memberships and the orthogonal decomposition used here. The two weightings answer different questions.

C.3 Alternative intervention games

A natural alternative to Equation (2) reconstructs the activation from the retained subset alone, $h(S) = \sum_{i \in S} \delta_i(x)$. This discards the reconstruction residual and the features outside the selected set. The full coalition then evaluates the behavior of the SAE reconstruction rather than the original model activation. The discarded component can also carry structure and behavioral effect (Engels et al., 2025b). Equation (2) keeps this component fixed, anchors the full coalition at h_x, and makes comparisons across SAE bases refer to the same original activation.

C.4 Scope of the zero point

The tested states lie in $h_x^0 + \mathrm{span}\{\delta_i(x) \mid i \in N_x\}$, whose dimension is at most K. A continuation can be nonlinear away from this subspace and still induce an additive response on the tested states. Behavioral compositionality therefore concerns the downstream computation expressed in the selected SAE coordinates. This is the coordinate space used when those contributions form a joint explanation or intervention.

D Experimental Specifications

D.1 Behavior Specifications

Table 6 lists the prompt sources and behavioral readouts used in the main experiments.

Table 6: Behavior definitions used in the main experiments.

Behavior	Prompt or source	Scalar behavior ψ
Affine control	Simple copying prompts	Linear readout of the intervened activation
Simple copying	Repeated name prompts	Target name logit difference
Subject verb agreement	Agreement prompts ending before the verb	Correct verb logit difference
Gendered pronoun preference	Biography prompts ending before the pronoun	Pronoun logit difference
CounterFact	Factual prompts from CounterFact (Meng et al., 2022)	Target object token logit
IOI	Indirect object identification prompts (Wang et al., 2023)	Indirect object logit difference
Greater than	Numerical comparison prompts	Correct answer label logit difference

Greater than uses answer labels that are compatible with each model tokenizer. Gemma-3 4B and Llama-3.1 8B use A and B labels. Mistral-7B uses True and False. The prompt templates and selected prompt identities are fixed across the corresponding comparisons.

D.2 STEERING SPECIFICATIONS

The steering experiment uses Llama-3.1 8B with the Llama Scope residual SAEs used by CorrSteer (Cho et al., 2026; He et al., 2024). We take the 16 MMLU-Pro features with the largest published positive correlations among the reported layerwise selections and keep their published steering coefficients fixed. Table 7 lists the resulting feature set.

Table 7: Fixed SAE features used in the MMLU-Pro steering experiment.

Index	Layer	Feature ID	Steering coefficient
1	26	22133	0.813
2	19	6432	0.966
3	22	11286	0.697
4	18	2624	0.478
5	20	28406	0.628
6	17	26604	0.469
7	24	7967	0.837
8	23	21146	0.853
9	25	16619	0.864
10	21	15538	0.391
11	15	9446	0.425
12	16	6219	0.342
13	14	28135	0.490
14	30	19427	0.374
15	12	28750	0.563
16	28	23202	1.120

A steering coalition S adds the selected direction at each corresponding layer with its fixed coefficient. We sample 1,024 distinct coalitions uniformly from the 2^{16} possible combinations and fix an equal fitting and evaluation split. The same coalition masks and split are used for every MMLU-Pro domain. We sample 100 test questions from each of the 14 official MMLU-Pro domains (Wang et al., 2024) and define the value of a coalition as its accuracy within that domain. Overall MMLU-Pro accuracy is the average over the resulting 1,400 questions. The reported confidence intervals resample prompts with replacement and recompute the accuracy games and projections.

The finite sample pairwise fit is evaluated on coalitions that were not used for fitting. It can therefore have a larger held out residual than the first order fit in some domains. We report $\widehat{\rho}_2$ without clipping. A negative value records a larger held out residual for the second order fit, while the population quantity ρ_2 remains nonnegative.

D.3 Model and SAE Pipeline Specifications

Table 8 summarizes the model, SAE family, and layers included in the analysis. The GPT-2 and Pythia SAEs follow the releases described by Bloom et al. (2024); the Gemma-3 SAEs use Gemma Scope 2 (McDougall et al., 2025), the Mistral SAEs use the residual SAEs of Engels et al. (2025a), and the Llama SAEs use Llama Scope (He et al., 2024).

Table 8: Model and SAE pipeline coverage.

Model	SAE family	Layers analyzed
GPT-2 Small	Bloom GPT-2 SAE	4, 6, 8, 10
Pythia-70M	Bloom Pythia SAE	1, 3, 5
Gemma-3 270M	Gemma Scope 2	5, 9, 12, 15
Gemma-3 1B	Gemma Scope 2	7, 13, 17, 22
Gemma-3 4B	Gemma Scope 2	9, 17, 22, 29
Mistral-7B	Engels residual SAE	8, 16, 24
Llama-3.1 8B	Llama Scope	8, 16, 21, 27

For each prompt, N_x contains the top K active features by activation magnitude. The released decoding convention of each SAE is used when converting feature activations to decoder contributions. The three main pipelines use unquantized model evaluation. Prompt identities, selected feature identities, coalition assignments, and fitting and evaluation splits are fixed within each comparison.

E Measurement Calibration

Table 9 reports synthetic checks of the residual estimator before the affine controls in the language models.

Table 9: Synthetic checks for the SAR estimator.

Game	$\widehat{\mathrm{SAR}}_1$	$\widehat{\mathrm{SAR}}_2$
Additive	6.1×10^{-12}	2.6×10^{-10}
Pairwise with target residual 0.1	0.100	≈ 0
Pairwise with target residual 0.3	0.300	≈ 0
Pairwise with target residual 0.5	0.500	≈ 0
Third order	0.180	0.066

Across the 11 affine controls in Gemma-3 4B, Mistral-7B, and Llama-3.1 8B, the largest prompt median $\widehat{\mathrm{SAR}}_1$ is 1.55×10^{-13} with 1,000 prompts for each control.

F Coordinate Comparisons

The same selected activation subspace can be written with different coordinate decompositions through Equation (5). For each prompt, we sample eight rotations independently by fixing the normalized all ones direction and QR orthogonalizing a Gaussian matrix on its $(K-1)$ dimensional orthogonal complement. Table 10 reports the main fixed layer comparison.

Table 10: Estimated first order SAR for learned SAE coordinates and random coordinate rotations within the same selected activation subspace. Each row uses 100 prompts and eight rotations per prompt. The final column gives the fraction of prompts whose SAE estimate lies below the median of its rotations.

Model	Behavior	SAE $\widehat{\mathrm{SAR}}_1$	Rotation $\widehat{\mathrm{SAR}}_1$	SAE below rotation median
Gemma-3 4B	Simple copying	0.04097	0.08333	0.98
	Subject verb agreement	0.03332	0.07638	0.99
	Gendered pronoun preference	0.03872	0.08080	1.00
	CounterFact	0.05558	0.09428	0.92
	IOI	0.02132	0.02402	0.66
	Greater than	0.02165	0.03198	0.98
Mistral-7B	Simple copying	0.02348	0.05794	0.98
	Subject verb agreement	0.02179	0.1196	1.00
	Gendered pronoun preference	0.02428	0.1411	0.98
	CounterFact	0.05545	0.1331	1.00
	IOI	0.01589	0.03812	1.00
	Greater than	0.06456	0.06556	0.54
Llama-3.1 8B	Simple copying	0.05932	0.1661	0.99
	Subject verb agreement	0.01944	0.07024	0.99
	Gendered pronoun preference	0.1089	0.1488	0.87
	CounterFact	0.08039	0.1377	0.90
	IOI	0.04875	0.07851	0.95
	Greater than	0.06296	0.1776	1.00

We also varied the intervention without changing the selected feature identities. Across nine representative model and behavior combinations, increasing feature inclusion from $p = 0.5$ to $p = 0.75$ preserves pairwise rank improvement in all nine combinations. Reducing each removed contribution to half of its observed magnitude lowers $\widehat{\mathrm{SAR}}_1$ in eight of nine combinations and increases first order rank fidelity in eight of nine. At smaller perturbation magnitude, the downstream response is therefore closer to a first order description.

G Full Behavioral Compositionality Results

G.1 Fixed Layer Behavior Comparisons

Table 11 gives the fixed layer estimates corresponding to the behavior comparison in the main text.

Table 11: Fixed layer behavior estimates with 1,000 prompts for each combination. Brackets give 95% confidence intervals from prompt bootstrap resampling.

Pipeline and layer	Behavior	n	Median $\widehat{\mathrm{SAR}}_1$	Median $\widehat{\rho}_2$
Gemma-3 4B L22	Simple copying	1000	0.03668 [0.03523, 0.03791]	0.6530 [0.6368, 0.6687]
	Subject verb agreement	1000	0.03732 [0.03496, 0.03877]	0.6274 [0.5978, 0.6518]
	Gendered pronoun preference	1000	0.03532 [0.03426, 0.03634]	0.8296 [0.8247, 0.8348]
	CounterFact	1000	0.05275 [0.05128, 0.05423]	0.7348 [0.7190, 0.7447]
	IOI	1000	0.04174 [0.03867, 0.04507]	0.7346 [0.7229, 0.7412]
	Greater than	1000	0.02121 [0.02090, 0.02162]	0.2814 [0.2653, 0.2949]
Mistral-7B L24	Simple copying	1000	0.02029 [0.01953, 0.02115]	0.7882 [0.7783, 0.7944]
	Subject verb agreement	1000	0.02435 [0.02304, 0.02538]	0.8718 [0.8503, 0.8842]
	Gendered pronoun preference	1000	0.04939 [0.04431, 0.05533]	0.8166 [0.8070, 0.8272]
	CounterFact	1000	0.05535 [0.05259, 0.05862]	0.8255 [0.8193, 0.8328]
	IOI	1000	0.01510 [0.01460, 0.01577]	0.7552 [0.7452, 0.7628]
	Greater than	1000	0.04159 [0.03366, 0.04756]	0.7804 [0.7757, 0.7854]
Llama-3.1 8B L21	Simple copying	1000	0.04486 [0.04313, 0.04698]	0.6813 [0.6739, 0.6885]
	Subject verb agreement	1000	0.03070 [0.02853, 0.03253]	0.7837 [0.7716, 0.7937]
	Gendered pronoun preference	1000	0.08244 [0.07991, 0.08794]	0.7778 [0.7709, 0.7864]
	CounterFact	1000	0.07707 [0.07241, 0.08102]	0.6855 [0.6713, 0.6998]
	IOI	1000	0.04245 [0.04046, 0.04401]	0.7152 [0.7078, 0.7204]
	Greater than	1000	0.05625 [0.05505, 0.05786]	0.7085 [0.6931, 0.7231]

G.2 Paired Early and Late Comparisons

Table 12 reports the paired endpoints used to summarize the depth profiles.

Table 12: Paired early to late estimates for the three detailed depth profiles. Brackets in the reduction column give 95% confidence intervals from paired prompt bootstrap resampling.

Pipeline	Behavior	Early $\to$ late $\widehat{\mathrm{SAR}}_1$	Reduction	Early $\to$ late $\widehat{\rho}_2$
Gemma-3 4B	Simple copying	0.1122 $\to$ 0.01155	9.711 [9.147, 10.29]	0.5835 $\to$ 0.7085
	CounterFact	0.1723 $\to$ 0.07646	2.254 [2.112, 2.483]	0.6929 $\to$ 0.8614
	IOI	0.2255 $\to$ 0.07320	3.080 [2.948, 3.248]	0.5615 $\to$ 0.8358
Mistral-7B	Simple copying	0.1771 $\to$ 0.02029	8.728 [8.232, 9.242]	0.5261 $\to$ 0.7882
	CounterFact	0.2658 $\to$ 0.05535	4.801 [4.432, 5.201]	0.4415 $\to$ 0.8255
	IOI	0.1270 $\to$ 0.01510	8.412 [7.865, 8.934]	0.3411 $\to$ 0.7552
Llama-3.1 8B	Simple copying	0.2001 $\to$ 0.2562	0.781 [0.737, 0.831]	0.5419 $\to$ 0.7782
	CounterFact	0.2817 $\to$ 0.09233	3.051 [2.872, 3.305]	0.4822 $\to$ 0.7916
	IOI	0.2151 $\to$ 0.1769	1.216 [1.123, 1.303]	0.5345 $\to$ 0.7813

G.3 Simple Copying Across Seven Pipelines

Figure 7 and Table 13 give the full comparison across seven pipelines referenced in the main text.

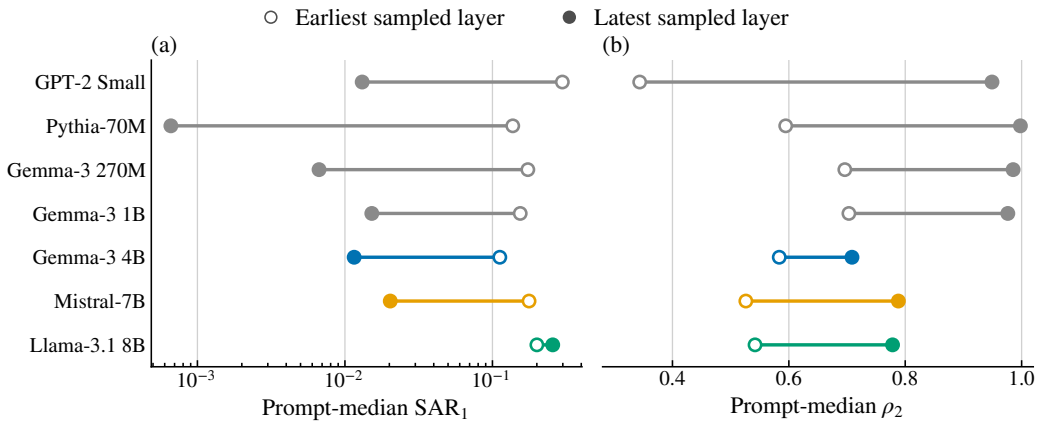


Figure 7: Simple copying at the earliest and latest sampled layers across seven model and SAE pipelines. Open markers denote the earliest layer and filled markers denote the latest layer. The left panel reports prompt median $\widehat{\mathrm{SAR}}_1$ and the right panel reports prompt median $\widehat{\rho}_2$.

Table 13: Estimated simple copying endpoints across seven model and SAE pipelines.

Pipeline	Early layer	Late layer	Early $\rightarrow$ late $\widehat{\mathrm{SAR}}_1$	Early $\rightarrow$ late $\widehat{\rho}_2$
GPT-2 Small	4	10	0.2978 $\rightarrow$ 0.01309	0.3437 $\rightarrow$ 0.9491
Pythia-70M	1	5	0.1373 $\rightarrow$ 0.0006608	0.5946 $\rightarrow$ 0.9979
Gemma-3 270M	5	15	0.1738 $\rightarrow$ 0.006676	0.6962 $\rightarrow$ 0.9854
Gemma-3 1B	7	22	0.1543 $\rightarrow$ 0.01521	0.7032 $\rightarrow$ 0.9763
Gemma-3 4B	9	29	0.1122 $\rightarrow$ 0.01155	0.5835 $\rightarrow$ 0.7085
Mistral-7B	8	24	0.1771 $\rightarrow$ 0.02029	0.5261 $\rightarrow$ 0.7882
Llama-3.1 8B	8	27	0.2001 $\rightarrow$ 0.2562	0.5419 $\rightarrow$ 0.7782

G.4 Greater Than Depth Profiles

Table 14 reports the complete greater than depth measurements available for Gemma and Llama.

Table 14: Estimated greater than depth profiles in the two pipelines with complete depth measurements.

Pipeline	Layer	Median $\widehat{\mathrm{SAR}}_1$	Median $\widehat{\mathrm{SAR}}_2$	Median $\widehat{\rho}_2$
Gemma-3 4B	9	0.2045	0.07585	0.6223
	17	0.1436	0.09039	0.3809
	22	0.02121	0.01495	0.2814
	29	0.06386	0.009314	0.8551
Llama-3.1 8B	8	0.3227	0.1005	0.6623
	16	0.1302	0.03133	0.7385
	21	0.05625	0.01509	0.7085
	27	0.2147	0.04577	0.7822

H Explanation Fidelity Results

For a held out coalition S, let

$$e(S) = g_x(S) - g_x(N_x), \qquad \widehat{e}_r(S) = \widehat{a}_r(S) - \widehat{a}_r(N_x).$$

Sign error is the fraction of evaluated coalitions for which $e(S)$ and $\widehat{e}_r(S)$ have opposite signs. For the same evaluation set $\mathcal{U}$, let $S^\star = \arg\max_{S\in\mathcal{U}} g_x(S)$ and $\widehat{S}_r = \arg\max_{S\in\mathcal{U}} \widehat{a}_r(S)$. Selection regret is

$$R^{(r)} = \frac{g_x(S^\star) - g_x(\widehat{S}_r)}{\max_{S\in\mathcal{U}} g_x(S) - \min_{S\in\mathcal{U}} g_x(S) + \epsilon}. \tag{48}$$

These measures supplement rank fidelity and behavior weighted ordering error.

H.1 Estimated SAR and First Order Fidelity

Table 15 gives the endpoint values for the $\widehat{\mathrm{SAR}}_1$ decile comparison in Figure 4.

Table 15: First order explanation fidelity in the lowest and highest $\widehat{\mathrm{SAR}}_1$ deciles within each model.

	Rank fidelity		Selection regret	
Model	Lowest decile	Highest decile	Lowest decile	Highest decile
Gemma-3 4B	0.9857	0.9539	0.01812	0.03623
Mistral-7B	0.9939	0.9413	0.009118	0.04684
Llama-3.1 8B	0.9790	0.9289	0.02711	0.05173

H.2 Pairwise Approximation

Tables 16 and 17 report the complete first order and pairwise results for the 18 model and behavior combinations.

Table 16: First order and pairwise explanation errors for all 18 model and behavior combinations.

		Sign error		Selection regret	
Model	Behavior	First order	Pairwise	First order	Pairwise
Gemma-3 4B	Simple copying	0.0892	0.0420	0.0150	0.0044
	Subject verb agreement	0.0784	0.0566	0.0175	0.0083
	Gendered pronoun preference	0.1070	0.0415	0.0261	0.0050
	CounterFact	0.0457	0.0418	0.0176	0.0034
	IOI	0.1275	0.0514	0.0288	0.0041
	Greater than	0.1288	0.1072	0.0269	0.0090
Mistral-7B	Simple copying	0.0451	0.0231	0.0185	0.0016
	Subject verb agreement	0.1199	0.0482	0.0328	0.0032
	Gendered pronoun preference	0.0864	0.0487	0.0301	0.0023
	CounterFact	0.0960	0.0684	0.0251	0.0037
	IOI	0.0474	0.0259	0.0101	0.0018
	Greater than	0.1231	0.0608	0.0350	0.0044
Llama-3.1 8B	Simple copying	0.0730	0.0433	0.0275	0.0052
	Subject verb agreement	0.0957	0.0500	0.0487	0.0077
	Gendered pronoun preference	0.0997	0.0612	0.0370	0.0109
	CounterFact	0.0861	0.0564	0.0378	0.0060
	IOI	0.0748	0.0423	0.0195	0.0036
	Greater than	0.0985	0.0484	0.0361	0.0065

Table 17 reports the corresponding rank fidelity values for the same model and behavior combinations.

Table 17: Mean prompt level rank fidelity of the first order and pairwise approximations for the 18 model and behavior combinations.

Model	Behavior	$F^{(1)}_{\text{rank}}$	$F^{(2)}_{\text{rank}}$
Gemma-3 4B	Simple copying	0.9857	0.9970
	Subject verb agreement	0.9883	0.9967
	Gendered pronoun preference	0.9690	0.9964
	CounterFact	0.9851	0.9961
	IOI	0.9751	0.9959
	Greater than	0.9744	0.9887
Mistral-7B	Simple copying	0.9912	0.9987
	Subject verb agreement	0.9745	0.9972
	Gendered pronoun preference	0.9654	0.9956
	CounterFact	0.9826	0.9973
	IOI	0.9930	0.9988
	Greater than	0.9897	0.9975
Llama-3.1 8B	Simple copying	0.9831	0.9963
	Subject verb agreement	0.9685	0.9961
	Gendered pronoun preference	0.9608	0.9942
	CounterFact	0.9669	0.9933
	IOI	0.9810	0.9968
	Greater than	0.9731	0.9943

H.3 Behavior Weighted Ordering Error

Table 18 reports the behavior weighted ordering errors used in the main text. Values use the original fitting and evaluation split.

Table 18: Held out behavior weighted ordering error for first order and pairwise approximations in the 18 model and behavior combinations used in the fidelity analysis.

Model	Behavior	$\widehat{\mathcal{E}}_1$	$\widehat{\mathcal{E}}_2$
Gemma-3 4B	Simple copying	0.0009103	0.0002062
	Subject verb agreement	0.0005988	0.0001896
	Gendered pronoun preference	0.001169	0.0001176
	CounterFact	0.0009869	0.0002584
	IOI	0.001248	0.0002133
	Greater than	0.0007911	0.0005375
Mistral-7B	Simple copying	0.0004209	0.00004362
	Subject verb agreement	0.0007353	0.00003234
	Gendered pronoun preference	0.001413	0.0001367
	CounterFact	0.0008927	0.0001262
	IOI	0.0002645	0.00003632
	Greater than	0.0004570	0.0001393
Llama-3.1 8B	Simple copying	0.001344	0.0002101
	Subject verb agreement	0.0009746	0.00009669
	Gendered pronoun preference	0.003579	0.0003866
	CounterFact	0.002580	0.0005645
	IOI	0.001357	0.0002247
	Greater than	0.002097	0.0003523

H.4 Pairwise Terms and Matched Higher Order Capacity

The pairwise model contains 137 coefficients. For the matched comparison, each control model keeps the intercept and 16 first order Walsh terms and replaces the 120 pairwise terms with 120 terms sampled uniformly without replacement from the 560 third order Walsh terms. Fifty control models are fitted for every prompt with the same fitting and evaluation coalitions as the pairwise model.

Table 19: Held out ordering for the pairwise model and parameter matched models that replace pairwise terms with randomly sampled third order Walsh terms. All columns report prompt medians; matched quantities first take the median over 50 random models for each prompt.

Model	Behavior	Pairwise rank	Matched rank	Pairwise $\widehat{\mathcal{E}}$	Matched $\widehat{\mathcal{E}}$
Gemma-3 4B	Simple copying	0.9934	0.9802	0.0002062	0.001488
	Subject verb agreement	0.9934	0.9845	0.0001896	0.0009812
	Gendered pronoun preference	0.9925	0.9597	0.0001176	0.001838
	CounterFact	0.9926	0.9772	0.0002584	0.001730
	IOI	0.9925	0.9690	0.0002133	0.001799
	Greater than	0.9800	0.9682	0.0005375	0.001151
Mistral-7B	Simple copying	0.9974	0.9885	0.00004362	0.0006113
	Subject verb agreement	0.9943	0.9674	0.00003234	0.001186
	Gendered pronoun preference	0.9912	0.9530	0.0001367	0.002338
	CounterFact	0.9946	0.9760	0.0001262	0.001579
	IOI	0.9976	0.9904	0.00003632	0.0004264
	Greater than	0.9949	0.9847	0.0001393	0.0007631
Llama-3.1 8B	Simple copying	0.9933	0.9766	0.0002101	0.001939
	Subject verb agreement	0.9924	0.9593	0.00009669	0.001486
	Gendered pronoun preference	0.9899	0.9517	0.0003866	0.005213
	CounterFact	0.9864	0.9578	0.0005645	0.003916
	IOI	0.9937	0.9756	0.0002247	0.002034
	Greater than	0.9898	0.9661	0.0003523	0.003020

H.5 Steering Results Across MMLU-Pro Domains

Table 20 gives the steering results for each domain summarized in the main text.

Table 20: Held out residual estimates and rank fidelity for steering combinations on MMLU-Pro. The first order and pairwise models are fitted on one set of sampled coalitions and evaluated on a separate set. The final column reports $F^{(2)}_{\mathrm{rank}} - F^{(1)}_{\mathrm{rank}}$.

Domain	$\widehat{\mathrm{SAR}}_1$	$F^{(1)}_{\mathrm{rank}}$	$\widehat{\mathrm{SAR}}_2$	$\widehat{\rho}_2$	$F^{(2)}_{\mathrm{rank}}$	Rank fidelity change
Biology	0.5815	0.6463	0.5600	0.0370	0.6684	+0.0221
Business	0.6666	0.5417	0.3878	0.4183	0.6784	+0.1367
Chemistry	0.9179	0.3170	0.5650	0.3844	0.7130	+0.3960
Computer Science	0.5420	0.6841	0.5356	0.0119	0.6962	+0.0121
Economics	0.5781	0.6558	0.6150	-0.0639	0.6465	-0.0093
Engineering	0.2238	0.8731	0.1871	0.1637	0.8689	-0.0042
Health	0.8685	0.3698	0.6871	0.2089	0.5922	+0.2224
History	0.6950	0.5716	0.4261	0.3869	0.7635	+0.1919
Law	0.7685	0.5062	0.9275	-0.2069	0.4155	-0.0907
Math	0.6151	0.6244	0.6543	-0.0638	0.5988	-0.0256
Other	0.4958	0.6889	0.4672	0.0577	0.6976	+0.0087
Philosophy	0.4549	0.7750	0.3637	0.2005	0.8039	+0.0289
Physics	0.9071	0.3133	0.9421	-0.0387	0.3669	+0.0536
Psychology	0.5055	0.6956	0.5085	-0.0060	0.7314	+0.0358
Overall	0.6424	0.5874	0.3891	0.3942	0.7790	+0.1916

Table 21: Held out behavior weighted ordering error for the steering responses.

Domain	$\widehat{\mathcal{E}}_1$	$\widehat{\mathcal{E}}_2$
Biology	0.1187	0.1080
Business	0.1287	0.06926
Chemistry	0.3078	0.09323
Computer Science	0.1043	0.09596
Economics	0.1071	0.1073
Engineering	0.02049	0.02116
Health	0.2634	0.1301
History	0.1520	0.06088
Law	0.1904	0.2429
Math	0.1243	0.1368
Other	0.07913	0.08483
Philosophy	0.06537	0.04759
Physics	0.2744	0.2465
Psychology	0.09043	0.07677
Overall	0.1444	0.06056

H.6 Reconstruction and Behavioral Compositionality Across Six SAEs

Figure 8 shows the comparison across the six SAE bases referenced in the main text.

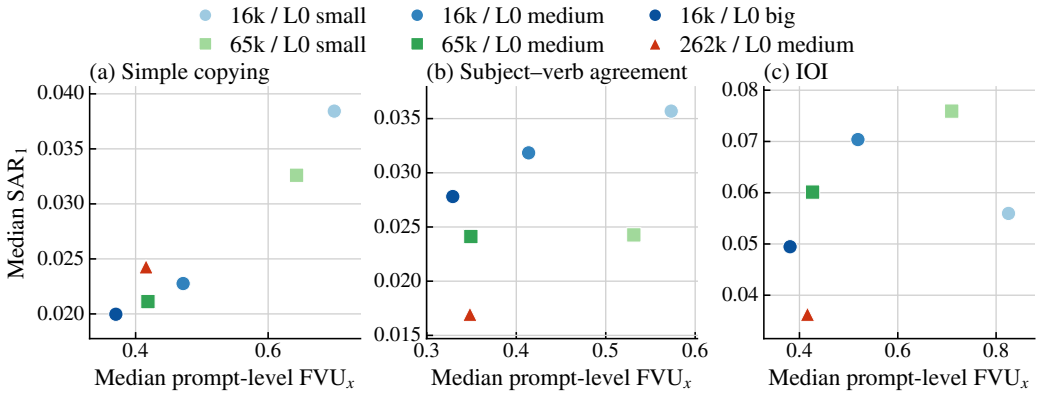


Figure 8: Median prompt level FVU_x and median $\widehat{\mathrm{SAR}}_1$ for six SAEs at Gemma-3 1B layer 17. Panels show simple copying, subject verb agreement, and IOI.

H.7 Examples

Figure 9 and Table 22 show the selected examples for individual prompts and their residual values.

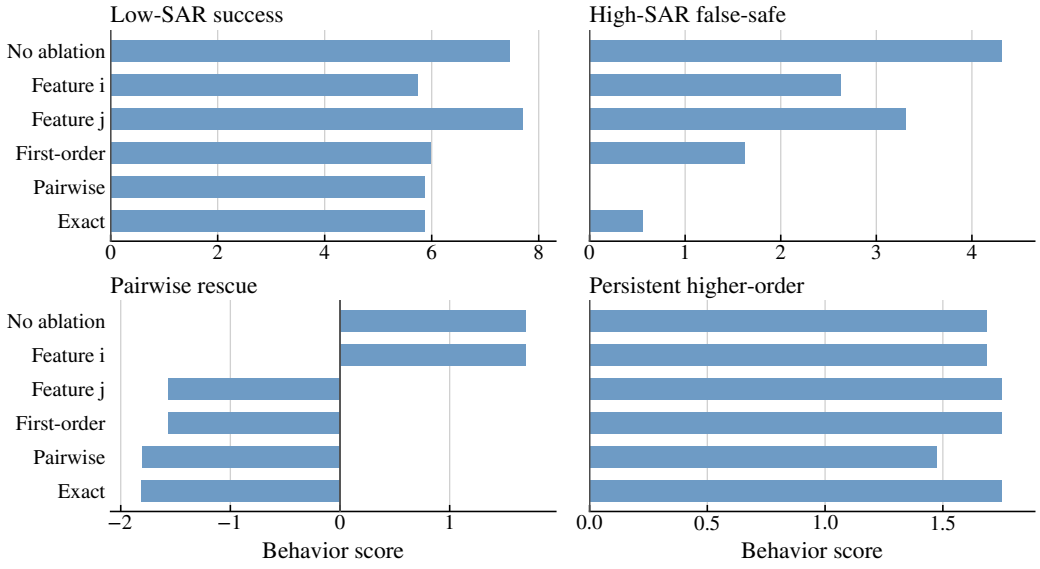


Figure 9: Behavior scores for four examples under ablations of individual features and feature pairs, together with first order and pairwise predictions when available.

Table 22: Prompts used in Figure 9.

Panel	Model	Behavior	Prompt ID	$\widehat{\mathrm{SAR}}_1$	$\widehat{\mathrm{SAR}}_2$	$\widehat{\rho}_2$
(a)	Mistral-7B	Simple copying	`simple_copying/02819`	0.00325	0.00159	0.509
(b)	Llama-3.1 8B	Gendered pronoun preference	`gendered_pronoun/00310`	0.1677	0.0384	0.771
(c)	Llama-3.1 8B	Subject verb agreement	`subject_verb_agreement/03544`	0.00765	0.00103	0.865
(d)	Llama-3.1 8B	Gendered pronoun preference	`gendered_pronoun/00072`	0.1412	0.0417	0.705

I Robustness and Protocol Checks

I.1 Sensitivity to the number of selected features

Changing K changes the candidate explanation by adding or removing SAE coordinates. The table therefore reports how behavioral compositionality changes with the explanation size rather than convergence of a fixed quantity.

Table 23: Sensitivity of $\widehat{\mathrm{SAR}}_1$ to the number of selected SAE features. The active feature rows report prompt median $\widehat{\mathrm{SAR}}_1$ at Gemma-3 1B layer 17. Affine control values are reported in units of 10^{-14}. The steering row reports $\widehat{\mathrm{SAR}}_1$ for the overall MMLU-Pro accuracy game. Feature sets are nested across K.

	Number of selected features K							
Behavior or setting	6	8	10	12	14	16	18	20
Affine control	6.87	6.40	5.70	5.16	5.03	5.43	5.07	4.84
Greater than	0.0460	0.0575	0.0634	0.0707	0.0763	0.0823	0.0868	0.0872
IOI	0.0554	0.0559	0.0597	0.0640	0.0652	0.0662	0.0664	0.0693
Simple copying	0.0213	0.0217	0.0234	0.0245	0.0270	0.0299	0.0307	0.0320
MMLU-Pro steering	0.2577	0.4951	0.6595	0.6350	0.5804	0.7768	0.8096	0.7861

Increasing the coalition inclusion probability from 0.5 to 0.65 preserves the decrease in $\widehat{\mathrm{SAR}}_1$ from the earliest to the latest sampled layer for CounterFact, IOI, and simple copying. Mean and median aggregation give the same depth directions. Using 1,000 prompts also preserves the main summary results, including decreases in $\widehat{\mathrm{SAR}}_1$ in eight of nine comparisons and increases in $\widehat{\rho}_2$ in all nine.