AUDITING BEHAVIORAL COVERAGE IN MODEL ECOSYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

A target model can differ from each of its peers yet be closely approximated by their fixed combination. A similarly small reconstruction error can also depend almost entirely on one peer. We study these distinct sources of behavioral coverage through matched model responses. The Peer-Inexpressible Residual (PIER) measures what remains after a fixed convex peer fit, and DISCO estimates the fit and its residual on separate samples. Comparing the group with a selected single peer, and refitting after peer removal, identifies who supplies the coverage. In an eight-model MMLU-Pro audit, our primary response is the reference-answer probability averaged over cyclic option rotations. Six targets benefit from collective approximation under two stress families and matched squared- and absolute-error fitting. Qwen2.5 instead obtains essentially all available approximation quality from its reasoning-tuned sibling; the reverse direction gains a further 15–18% from other peers. Mistral's group fit reduces error by 19–20%. A separate complete-option-distribution analysis retains the Qwen directionality and Mistral's group advantage. Matched stress increases residuals for some targets and decreases them for others. In vision, context changes both the supporting peer combination and the concentration of residuals on individual images. These results characterize how accurately a collection reproduces a target, which peers supply that approximation, and which inputs account for the remaining gaps.

Code: https://anonymous.4open.science/r/behavioral-coverage-audit-F68D

1 INTRODUCTION

Two models can be close to one another and still play different roles in a larger collection. One may be reconstructed almost as well by a closely related peer as by all available peers, while reconstructing that related model benefits from additional models. Conversely, a target can differ from every individual peer yet be closely approximated by their combination. The same pairwise comparison cannot distinguish dependence on one peer from dependence on the group. This distinction concerns *who covers whom*: how the collection reproduces a target's responses, and which peers make that reproduction possible.

A reconstruction score alone also leaves this structure unresolved. Two targets with similar residuals may rely on very different subsets of the ecosystem. Removing a single related checkpoint can disrupt the approximation of one target while scarcely affecting the other. The relevant comparison therefore combines reconstruction error with the change in error when the available peers change. It gives an auditor a target-specific account of where the collection's coverage comes from.

We formalize *behavioral coverage* through recorded responses on matched inputs. An audit specifies the response to observe, the peer set, and the fitting and evaluation conditions. A fixed convex combination uses the same non-negative, sum-to-one weights on every audited input. This asks whether the peers' existing response traces jointly reproduce the target on their original scale. The Peer-Inexpressible Residual (PIER) is the remaining target–peer difference. DISCO fits the combination on one sample and evaluates residual magnitude on another. Group gain compares that fit with a single peer selected on the fitting sample; peer-removal effects measure the loss of coverage after each peer is excluded and the combination refitted.

Our principal experiment studies eight instruction- and reasoning-tuned language models on MMLU-Pro (Wang et al., 2024). The primary response is each model's reference-answer probability, balanced

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.

over cyclic option rotations. One weight vector must reconstruct its variation across questions and stress conditions. Six targets benefit from a peer group under both stress families and both matched fitting losses. Qwen2.5 has a different structure: General Reasoner, its reasoning-tuned sibling, achieves essentially the full group's approximation quality. In the reverse direction, the Qwen sibling remains important, but additional peers reduce General Reasoner's error by 15–18%. Mistral exhibits a 19–20% group advantage. The weak removal effects within the Phi pair provide a contrast: lineage identifies candidate relations, while measured responses determine their strength.

We examine this structure through a second response view that retains the probabilities of *all* answer options. In the canonical-order vector analysis, the Qwen pair again differs in whether additional peers improve reconstruction, and Mistral retains a collective advantage. Gemma, however, has different group-gain signs in the two settings. The response map thus selects a substantive object of comparison: agreement in reference-answer confidence and agreement across the complete option distribution can reveal different coverage relations.

Matched stress adds a further dimension. Irrelevant context increases the balanced residual for Qwen2.5 and phi-4 while decreasing it for Mistral and OLMo. The vision experiments connect such changes to the supporting peer combination and to a small set of shared high-residual images. Forecasting data illustrate the separate role of task outcomes: a response difference can accompany either an improvement or a deterioration in prediction error. Together, these analyses distinguish the amount, source, and location of a collection's coverage.

The estimation step builds on convex response fitting, as in synthetic control (Abadie et al., 2010). Here the observations are matched model responses, and the object of study is their reconstruction relation. This differs from explanations of a single model (Ribeiro et al., 2016; Sundararajan et al., 2017; Wachter et al., 2017), comparisons of internal representations (Raghu et al., 2017; Kornblith et al., 2019), and combinations optimized for task prediction (Hinton et al., 2015; Wortsman et al., 2022).

Contributions.

- *Collective response coverage.* We formulate target reconstruction by a fixed peer-response class and combine held-out PIER, group gain, and removal effects to distinguish the amount of coverage from its source. Projection properties support the estimator even when coefficient representations are non-unique.
- *Directional and distributed relations.* An eight-model language-model audit separates single-sibling coverage from sibling-supported and distributed approximation. A complete-option-distribution diagnostic examines which of these patterns extend beyond reference-answer confidence.
- *Stress and localization.* Matched interventions move target residuals in opposite directions within one ecosystem. Response balancing tests those directions, and vision audits identify changes in peer support and concentrated shared residuals.

2 Behavioral coverage and peer expressibility

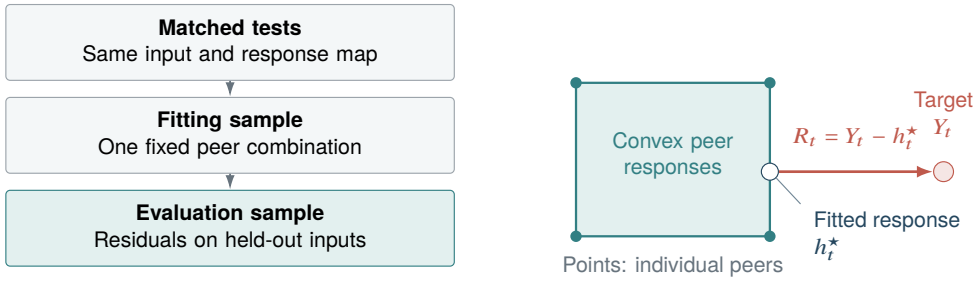
Let $\mathcal{J} = \{1, \dots, N\}$ index a model collection, t a target, and $\mathcal{J}_{-t}$ its peers. An audit observes the models on the same transformed input $T_{\theta,\omega}(x)$, where θ specifies an intervention level and ω its realization. Writing $z = (x, \theta, \omega)$, the recorded response is

$$Y_j(z) = g_x(f_j(T_{\theta,\omega}(x))) \in \mathbb{R}^q. \tag{1}$$

The response map g_x selects what the audit compares. For multiple-choice questions, it can record reference-answer confidence or the entire option-probability vector. For forecasting, it records the prediction. Model-specific wrappers preserve the same semantic input and recorded quantity across peers.

Fixed collective reconstruction. Let $\Delta^{N-2} = \{w \in \mathbb{R}^{N-1} : w_j \geq 0, \sum_j w_j = 1\}$. The peer-response class is

$$h_w(z) = \sum_{j \in \mathcal{J}_{-t}} w_j Y_j(z), \qquad \mathcal{C}_t = \{h_w : w \in \Delta^{N-2}\}. \tag{2}$$



(a) Matched responses and a fixed peer fit. **(b)** Projection onto the peer-response class.

Figure 1: A fixed peer fit reconstructs a response trace. A fixed combination is learned from matched fitting observations and evaluated on held-out inputs. Its residual records what remains unexplained by the peer class; comparing alternative peer sets identifies the source of coverage. Both panels are conceptual.

A single w applies across the audit. Even when $Y_j(z)$ is scalar, the fit reconstructs a whole response trace $(Y_j(z_1), \ldots, Y_j(z_m))$, rather than choosing coefficients separately for each observation.

Convex weights combine the existing responses on their original scale and preserve probability normalization for vector responses. Allowing signed coefficients would permit extrapolation; allowing input-dependent weights would permit adaptive selection. The fixed convex class isolates the coverage supplied by an unchanged combination of peers.

A fitting distribution P defines the population approximation,

$$h_t^\star \in \arg\min_{h \in \mathcal{C}_t} \mathbb{E}_P \|Y_t - h\|_2^2. \tag{3}$$

For evaluation distribution Q, the residual and its mean magnitude are

$$R_t(z) = Y_t(z) - h_t^\star(z), \qquad U_t(P, Q) = \mathbb{E}_Q \|R_t(Z)\|_2. \tag{4}$$

We call R_t PIER and U_t the *coverage gap*. Scalar responses use absolute residuals; the full-option analysis additionally reports total variation. Section 3 specifies the support conditions under which the fitted projection defines a unique evaluation response.

Estimation and peer dependence. DISCO solves the empirical simplex-constrained version of Eq. (3) on a fitting sample and evaluates $\widehat{U}_t$ on held-out observations. The audit requires model responses but no model parameters or gradients. Absolute-loss fitting supplies a second matched comparison in the scalar experiments; Section 3's projection uniqueness result applies to squared loss.

Let E_t^{single} be the held-out error of the peer selected on the fitting sample, and E_t^{group} the error of the fitted combination. Define

$$\Gamma_t = E_t^{\text{single}} - E_t^{\text{group}}, \tag{5}$$

$$I_{t \leftarrow j} = \widehat{U}_t(\mathcal{J}_{-t} \setminus \{j\}) - \widehat{U}_t(\mathcal{J}_{-t}). \tag{6}$$

Each reduced peer set is refitted under the same design. Positive Γ_t identifies a collective advantage; positive $I_{t \leftarrow j}$ measures how much coverage is lost without peer j. Together they separate dependence on one peer from improvement supplied by the rest. They are held-out comparisons and can have either sign.

3 Estimation and Representation

Two properties connect the empirical fit to a population response and determine how its coefficients should be interpreted.

Proposition 1 (Projection and representation)**.** *If the finitely many target and peer responses belong to $L^2(P; \mathbb{R}^q)$, then $\mathcal{C}_t$ is compact and convex in $L^2(P)$. Equation (3) has a unique projected response, although its representing weight vector may be non-unique. Enlarging the peer set cannot increase the optimized squared distance under the same P.*

Table 1: Exact-response controls. Mean and maximum held-out absolute residuals from cached probability responses; the first three targets are constructed controls.

Control	Mean residual	Maximum residual
Exact clone	2.71×10^{-11}	1.33×10^{-10}
Mixture $(0.5, 0.3, 0.2)$	2.96×10^{-11}	1.26×10^{-10}
Boundary mixture $(0.75, 0.25)$	2.55×10^{-11}	1.15×10^{-10}
Duplicated peer	2.48×10^{-11}	1.17×10^{-10}

The simplex has a compact image in response space. Projection onto that image is unique up to P-almost-sure equality. Coincident peers can exchange coefficient mass while preserving the response, so removal effects compare the coverage available from different peer sets rather than the size of a chosen coefficient.

Proposition 2 (Held-out consistency)**.** *Assume bounded responses, a fixed finite peer set, independent fitting and evaluation samples, and $Q \ll P$ with bounded density ratio. An empirical squared-loss minimizer satisfies $\|\widehat{h}_t - h_t^\star\|_{L^2(P)} \to_p 0$ and its held-out residual estimate satisfies $\widehat{U}_t \to_p U_t(P, Q)$ as both sample sizes increase.*

Uniform convergence over the simplex gives consistency of the fitting response; support compatibility transfers it to evaluation. The matched LLM half-splits use the same population design, with questions as the independent sampling units. Appendix A gives the proofs, the extension to clustered intervention realizations, and the distinction between fitting-loss monotonicity and held-out comparisons.

Exact-response controls. A copied response, a known three-peer mixture, and a boundary mixture give held-out mean residuals of order 10^{-11}. A duplicated peer changes the coefficient representation while preserving the fitted response (Table 1). These cases connect the implementation directly to the reconstruction and representation properties above.

4 Coverage structure in language-model ecosystems

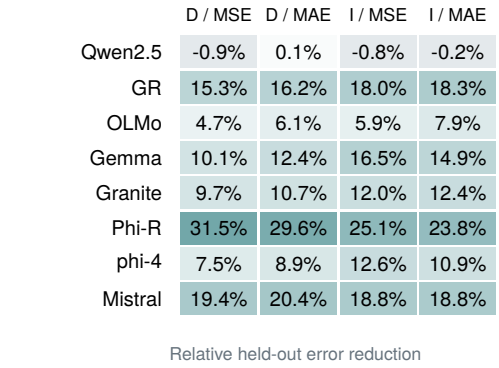
The residual describes the quality of an approximation; group gains and removal effects describe its source. We examine their joint structure in eight instruction- and reasoning-tuned checkpoints from the Qwen, Phi, Mistral, OLMo, Granite, and Gemma families. The Qwen and Phi sibling comparisons follow checkpoint lineage; the balanced-interface analysis evaluates these specified pairs. Appendix B gives checkpoint identities and the complete design.

Reference-answer response and matched fitting. The audit contains 560 MMLU-Pro questions across 14 subjects (Wang et al., 2024). Candidate labels are scored by conditional sequence likelihood and normalized over the available options. Every cyclic rotation is mapped back to semantic option identities before averaging; the primary response is the resulting reference-answer probability. Each option therefore occupies each label position once. This response tracks the target's reference-answer confidence across questions and interventions.

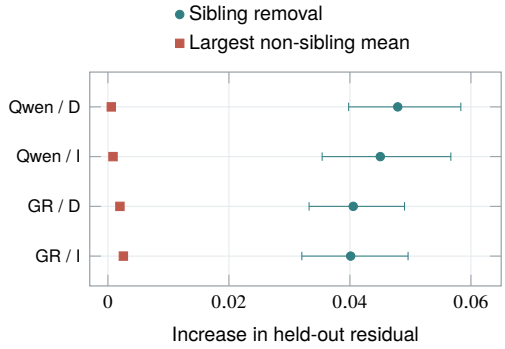
We use ten stratified fitting/evaluation half-splits. Each family assigns equal total mass to clean and maximum-stress conditions; three matched realizations share the stressed mass. Residual magnitudes are calculated for each realization before averaging. Single-peer selection and group fitting use the same loss, either mean squared error (MSE) or mean absolute error (MAE), with held-out MAE as the common evaluation. Pointwise 95% intervals use 1,000 common-question bootstrap draws: each physical question is resampled jointly across models and splits, and every fit is recomputed. This preserves the repeated splits' dependence.

4.1 Collective approximation varies across targets

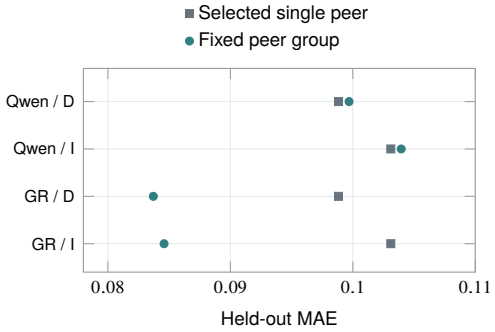
Six targets show positive group gains under both stress families and both fitting losses, with each corresponding pointwise interval excluding zero (Figure 2a). Relative improvements range from about 7% to 32%. OLMo has smaller gains, including a deletion result whose MSE-fit interval crosses zero. Qwen2.5 has no reliable group advantage. The all-target view therefore separates



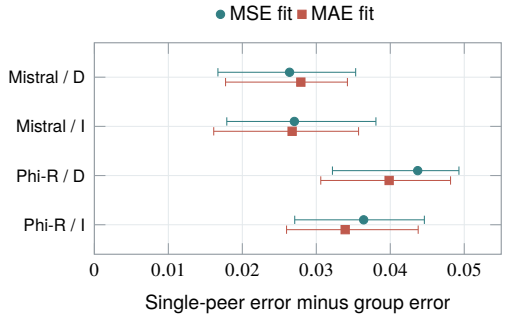
(a) Relative group gain across the ecosystem.



(b) Qwen sibling and non-sibling removal effects.



(c) Single-peer and group errors in both directions.



(d) Distributed gains under matched fitting losses.

Figure 2: A peer group can cover a target without a single peer doing so. All panels use permutation-averaged semantic probabilities. D/I denote deletion and irrelevant context; MSE/MAE denote the fitting loss. Panel (a) reports $100\Gamma_t/E_t^{\text{single}}$. Panel (b) compares declared sibling removal with the largest mean non-sibling effect. Panel (c) uses MSE fitting and held-out MAE. Panel (d) reports absolute group gains with pointwise 95% common-question bootstrap intervals. Phi-R denotes Phi-4-reasoning-plus; GR denotes General Reasoner.

recurrent collective approximation from a different regime in which one peer already achieves the group's available reconstruction quality.

4.2 A CLOSE PAIR HAS UNEQUAL ROLES IN THE COLLECTION

Removing General Reasoner raises Qwen2.5's gap by 0.04790 under deletion and 0.04502 under irrelevant context. The largest mean non-sibling removal effects are 0.00056 and 0.00084. In the reverse direction, removing Qwen2.5 raises General Reasoner's gap by 0.04055 and 0.04012. For each target and family, the declared sibling is the most consequential removal in all ten splits (Figure 2b).

These strong reciprocal removal effects coexist with a directional group advantage. General Reasoner alone achieves essentially the full group's error for the Qwen2.5 target. For the General Reasoner target, Qwen2.5 is still the central peer, but adding the rest of the collection lowers error by 15–18% (Figure 2c). The pairwise response difference is symmetric; the gain supplied by the remaining peers is not. Qwen2.5 thus exhibits *single-sibling coverage* relative to what the group can attain, while General Reasoner exhibits *sibling-supported collective coverage*.

The Phi pair separates shared lineage from strong reconstruction dependence. Removing phi-4 raises Phi-4-reasoning-plus's gap by 0.01119 under deletion and 0.00835 under irrelevant context, compared with largest mean non-sibling effects of 0.00577 and 0.00691. The sibling contributes, but its separation from other peers is much smaller than in the Qwen pair. In the reverse direction, removing Phi-4-reasoning-plus changes phi-4's irrelevant-context gap by -0.00161, with a pointwise interval spanning zero (Appendix C). Shared lineage therefore does not determine the strength or direction of measured dependence. The group comparison and the removal comparison supply distinct evidence: the first asks whether additional peers improve reconstruction; the second asks whether the remaining group can compensate for a particular peer's absence.

4.3 Distributed gains persist across fitting losses

Mistral's combination improves on the selected single peer by 19.4% under deletion with MSE fitting and 20.4% with MAE fitting. Under irrelevant context, both reductions are about 18.8%. The four intervals exclude zero (Figure 2d). Phi-4-reasoning-plus has larger reference-answer gains of approximately 24–32%.

For phi-4 itself, relative reductions are 7.5–8.9% under deletion and 10.9–12.6% under irrelevant context (Appendix C). Both Phi targets benefit from the collection despite weaker sibling-removal effects than the Qwen pair. Collective improvement and dependence on a particular sibling are therefore distinct aspects of coverage.

4.4 Complete option distributions retain the directional pattern

Reference-answer confidence and the allocation of probability among alternative answers are different response views. The complete-option audit fits all option probabilities jointly, using one coefficient vector across questions, three tracks, and five equally weighted doses. It uses the canonical answer order, vector squared-loss group fitting, and a single peer selected by fitting total variation (TV). Table 2 reports held-out TV from the completed ten-split diagnostic.

Both Qwen targets select their sibling in every split. Their mean pairwise TV is therefore identical: 0.314 under deletion and 0.337 under irrelevant context. Adding the other peers lowers General Reasoner's TV to 0.299 and 0.305, whereas the fitted group has higher TV for Qwen2.5, 0.342 and 0.363. Each direction holds across all ten splits. Mistral also retains a group advantage, with reductions of 14.3% and 15.4%. The two response views thus agree on which Qwen target benefits from additional peers and on Mistral's collective coverage.

Table 2: Coverage across the complete option distribution. Mean held-out total variation over ten splits in the canonical-order, five-dose audit. The group minimizes vector squared error; the single peer is selected by fitting total variation. This response diagnostic uses a different design from the balanced, loss-matched results in Figure 2. All models and removal effects appear in Appendix C.3.

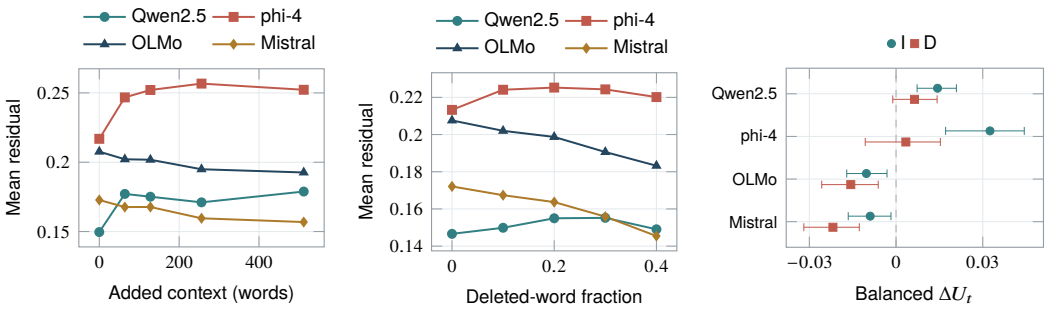
	Deletion		Irrelevant context	
Target	Single	Group	Single	Group
Qwen2.5	0.314	0.342	0.337	0.363
General Reasoner	0.314	0.299	0.337	0.305
Mistral	0.421	0.360	0.416	0.352
Gemma	0.486	0.543	0.468	0.538

Vector removal fits show that the Qwen sibling remains important even when group gain is absent (Table 3). Removing either sibling increases the other's held-out TV in every split and both families. Thus Qwen2.5's negative vector group gain does not mean that its peers are irrelevant: its sibling is a better reference than the fitted combination, yet removing that sibling leaves a substantially worse combination. The Phi comparison again differs. Removing phi-4 raises Phi-R's TV, whereas removing Phi-R lowers phi-4's TV under irrelevant context in all ten splits. These paired comparisons locate dependence beyond the reference-answer coordinate.

Table 3: Sibling-removal effects for complete option vectors. Increase in held-out TV after removing the declared sibling and refitting the vector squared-loss combination; means over ten fixed splits. The last column counts positive changes in the overlapping deletion/context splits; the full diagnostic is specified in Appendix C.3.

Target	Removed peer	Deletion	Context	Positive splits (D/I)
Qwen2.5	General Reasoner	+0.1365	+0.1172	10/10, 10/10
General Reasoner	Qwen2.5	+0.1076	+0.0856	10/10, 10/10
Phi-R	phi-4	+0.0203	+0.0162	10/10, 10/10
phi-4	Phi-R	+0.0030	-0.0087	9/10, 0/10

The agreement is selective. Gemma benefits in the balanced reference-answer audit but has higher group TV than single-peer TV in this canonical vector diagnostic. The protocol and fitting-loss differences are specified in the table: the comparison establishes coverage under two recorded views rather than isolating response dimension alone. A coefficient-transfer comparison further separates the response measurement from the choice of fit. Keeping the canonical scalar-fit weights fixed while evaluating complete option distributions gives General Reasoner mean TV of 0.2935 under deletion



(a) Canonical context trajectories. **(b)** Canonical deletion trajectories. **(c)** Balanced endpoint changes.

Figure 3: Controlled stress does not uniformly increase the coverage gap. Curves use fixed canonical-interface fits across five doses. Endpoint changes use separately fitted balanced-interface endpoint designs, with pointwise 95% common-question bootstrap intervals. Positive changes indicate greater reconstruction error. D denotes deletion; I denotes irrelevant context.

and 0.3009 under irrelevant context; for Mistral, the values are 0.3630 and 0.3557. Each is below its corresponding selected-single value in Table 2. Their collective advantage is therefore present both when the coefficients are learned from reference-answer responses and when all option coordinates enter fitting. Appendix C.3 gives the full transferred and refitted results for every target, including the cases where these views disagree.

5 Coverage under stress and response balancing

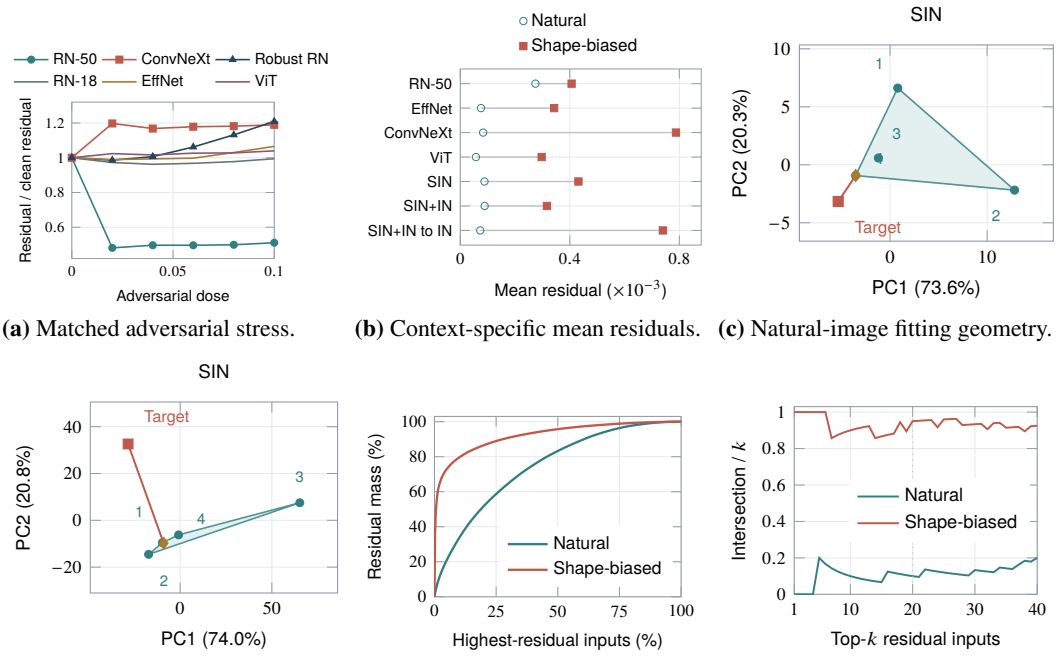
The identity of the peers is fixed, but the response relation can change with the input. Irrelevant-context injection prepends unrelated material; content deletion removes selected words from the question stem while preserving the options. A single combination is fitted across the declared conditions of each family and held fixed during evaluation. Changes in its residual reveal where that combination ceases to track the target.

Opposite directions under matched stress. At the maximum irrelevant-context dose, the balanced reference-answer gap increases by 0.01435 for Qwen2.5 and 0.03250 for phi-4. It decreases by 0.00894 for Mistral and 0.01026 for OLMo. All four directions hold in every split and have common-question bootstrap intervals excluding zero (Figure 3c). Under deletion, Mistral and OLMo also decrease, by 0.02183 and 0.01565. The same perturbation therefore makes one target harder to reconstruct and another easier.

The canonical-order curves show the full dose dependence (Figure 3a–b). Balanced endpoint fits then examine the high-dose changes under a second presentation protocol. The curves and endpoint comparisons retain their own response definitions and fitting designs. The complete-option diagnostic also gives opposite irrelevant-context directions for Qwen2.5 and Mistral (Appendix C.3).

Presentation changes magnitudes more often than directions. The mean rotation-to-average TV diagnostic is 0.338, indicating substantial variation in semantic option probabilities across presentations. Nevertheless, canonical and balanced endpoint analyses agree on 15 of 16 effect signs. Their eight-target rankings of family-averaged reconstruction errors have Spearman correlation 0.952; the correlation across the 16 endpoint changes is 0.932. These are complementary summaries of the error level and its response to stress.

Granite illustrates why both the effect and its uncertainty matter. Its canonical contractions remain negative after balancing, but the balanced intervals include zero. The resulting evidence supports the Qwen2.5, phi-4, Mistral, and OLMo irrelevant-context directions while distinguishing them from the weaker Granite effects. Thus presentation balancing refines which coverage changes are supported, rather than leaving every conclusion unchanged. Appendix C gives all endpoint estimates and the separate direct-generation diagnostics.



(a) Matched adversarial stress. **(b)** Context-specific mean residuals. **(c)** Natural-image fitting geometry.

(d) Shape-biased fitting geometry. **(e)** ConvNeXt residual concentration. **(f)** Shared highest-residual inputs.

Figure 4: Context affects residual size, peer support, and localization. Panels (a)–(b) summarize held-out responses; lines connect the observed dose levels. Panels (c)–(d) show one SIN target with separate PCA bases (93.9% and 94.8% explained variation). The mixture is projected from the full response space. Peer labels: 1 ResNet-50, 2 EfficientNet, 3 ConvNeXt, 4 ViT. Panels (e)–(f) use all 400 stored residuals per context; overlap compares ConvNeXt-T with the SIN+IN→IN ResNet in the seven-model audit.

6 Context-Dependent Gaps and Image-Level Residuals

An increased average residual can arise from many small discrepancies or from a few large ones. The vision audit distinguishes these patterns and relates them to the supporting peer combination. It uses ImageNet classifiers and shape-trained ResNet variants on matched images (Deng et al., 2009; He et al., 2016; Geirhos et al., 2019). Fitting and evaluation are separated within each context; the adversarial and context experiments retain their respective score interfaces (Appendix D).

Context changes the amount and source of coverage. At the maximum matched adversarial dose, ResNet-50's mean residual is 0.511 times its clean value, while ConvNeXt-T and robust ResNet-50 reach 1.189 and 1.209 (Figure 4a). The separate seven-model, true-class-probability audit shows higher context-specific gaps for all seven targets on shape-biased images. ConvNeXt-T increases from 8.35×10^{-5} to 7.88×10^{-4}, a 9.44-fold change (Figure 4b).

For the SIN-trained ResNet target with four standard peers, context-specific PCA views of the fitting response vectors show a concurrent change in support (Figure 4c–d). The natural-image fit is approximately the ViT response. The shape-biased fit instead assigns weights of about 0.507, 0.223, and 0.270 to ResNet-50, EfficientNet-B0, and ViT-B/16. Its full fitting-vector residual norm rises from 0.00488 to 0.04673. These quantities describe the original-space fit; each panel has its own PCA basis. Context changes both the target's discrepancy and which combination lies closest to it.

The other shape-trained targets show different support changes under the same four-peer roster (Appendix D). The SIN+IN target changes from approximately the EfficientNet response on natural images to a mixture dominated by ViT on shape-biased images. The SIN+IN→IN target changes from approximately ViT alone to weights of 0.813 on ConvNeXt and 0.187 on ViT. Their fitting-vector residual norms increase from 0.01076 to 0.02076 and from 0.00604 to 0.03309, respectively. The larger gap is therefore accompanied by a target-specific change in which peers provide the closest recorded approximation, rather than a uniform change in one shared reference model.

The gap concentrates on shared inputs. Among the 400 evaluation residuals per context, the largest 5% for ConvNeXt-T account for 72.99% of residual mass on shape-biased images and 20.98% on natural images (Figure 4e). ConvNeXt-T and the SIN+IN→IN ResNet share 19 of their 20 highest-residual positions in the shifted context, compared with two in the natural context (Figure 4f). These sample-level summaries identify a shared concentration of reconstruction difficulty. The residual field therefore adds localization to the aggregate gap: it distinguishes broad disagreement from a small subset of inputs that dominate the measured error.

7 Coverage and task utility

Reconstruction evaluates a peer response against the target; task loss evaluates it against the outcome. The 31-city forecasting archive (Loder et al., 2019) illustrates the distinction. Relative residual is $100\,\mathrm{PIER}/\mathrm{MeanFlow}$, and replacement impact is $100(\mathrm{MAE}_{\mathrm{peer}} - \mathrm{MAE}_{\mathrm{local}})/\mathrm{MAE}_{\mathrm{local}}$. Paris has a positive impact, while Bern has a negative one despite its large relative residual. The normalized quantities have Pearson correlation 0.052; the raw quantities have correlation 0.836. The definition of scale therefore matters, and response distinctiveness alone does not determine the sign of task-loss change.

We use this archive as a descriptive comparison of two measurement targets. Its stored fits include numerically infeasible simplex coefficients, so its complete point plots and coefficient diagnostics are retained in Appendix D.4, separately from the matched LLM comparisons. The empirical role of the coverage audit is to locate response dependence: the Qwen removal analysis identifies which peer supplies the available reconstruction, while task-outcome measurements assess the value of retaining or replacing that response.

8 Related work

Attribution and model comparison. Feature attribution allocates a model response to input variables, while counterfactual explanation studies changes that alter a decision (Ribeiro et al., 2016; Sundararajan et al., 2017; Wachter et al., 2017). SVCCA and CKA compare learned representations (Raghu et al., 2017; Kornblith et al., 2019). Our observations are output response traces: a target is reconstructed by a fixed peer combination, and alternative peer sets reveal its reconstruction dependence. Input-level residuals then connect the collection-level fit to the examples on which it fails.

Aggregation, transfer, and credit. Ensembles and model soups combine models to improve task prediction; distillation transfers behavior to a student (Hinton et al., 2015; Wortsman et al., 2022). The objective here is target-response reconstruction, so approximation error is a result rather than a training surrogate for task performance. Fixed convex fitting follows the synthetic-control construction (Abadie et al., 2010); matched interventions provide common observations for the model collection. Shapley values allocate a chosen coalition utility (Shapley et al., 1953; Lundberg and Lee, 2017). Their object differs from coverage: duplicated responses can receive positive credit while leaving zero peer residual. Appendix E formalizes this example and the support requirements of observational audits.

9 Discussion

Coverage has both an amount and a source. Qwen2.5 reaches its group's reference-answer approximation with one sibling; General Reasoner benefits from additional peers. Mistral retains collective gains across both response views. Stress changes residuals in opposite directions, while image-level residuals localize the gaps. The response map and fixed convex class determine what is reconstructed; open-ended generation and downstream ensemble value remain separate questions. Signed or input-dependent combinations would measure different forms of expressibility. Across the studied settings, residuals and removal effects distinguish how well a target is reconstructed from which peers make that reconstruction possible. Model dependence is thus an observable relation within a collection, rather than a label attached to one model.

USE OF AI STATEMENT

Large language models were used as assistive tools during this work. They were used to improve the clarity and language of author-written text, to support literature retrieval and discovery of related work, to assist with research ideation and experimental execution (including discussing experimental designs, generating and debugging research code, and analyzing intermediate results), and to assist in developing and checking mathematical arguments and proofs.

All research questions, methodological choices, experimental protocols, interpretations, mathematical claims, and conclusions were determined and critically reviewed by the authors. All AI-assisted code, references, calculations, proofs, and factual claims were independently inspected and verified by the authors. The authors take full responsibility for the scientific content and the final manuscript.

REPRODUCIBILITY STATEMENT

The source release includes standalone LaTeX panels, their numerical inputs, arithmetic replay scripts, and a provenance manifest. Appendix B specifies the primary LLM protocol and clustered bootstrap; Appendix C.3 gives the separate canonical complete-option analysis. Appendix D records the cross-domain observation interfaces and limitations of the stored numerical fits.

REFERENCES

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, et al. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

IBM Granite Team. Granite 3.3 8b instruct model card. https://huggingface.co/ibm-granite/granite-3.3-8b-instruct, 2025.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019.

Allister Loder, Lukas Ambühl, Monica Menendez, and Kay W Axhausen. Understanding traffic capacity of urban networks. *Scientific reports*, 9(1):16283, 2019.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Xueguang Ma, Qian Liu, Dongfu Jiang, Ge Zhang, Zejun Ma, and Wenhu Chen. General-reasoner: Advancing llm reasoning across all domains. *URL https://arxiv. org/abs/2505.14652*, 2025.

Mistral AI. Mistral nemo instruct model card. https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407, 2024.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Lloyd S Shapley et al. A value for n-person games. 1953.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. Pmlr, 2022.

A Projection theory and statistical scope

A.1 Existence and response uniqueness

Fix the fitting distribution P and write $\mathcal{H} = L^2(P; \mathbb{R}^q)$. Assume finitely many responses $Y_j \in \mathcal{H}$. The map $A : w \mapsto \sum_j w_j Y_j$ is continuous from the compact simplex into $\mathcal{H}$, so its image $\mathcal{C}_t$ is compact. It is also convex. The metric projection of Y_t onto this nonempty closed convex set exists and is unique in $\mathcal{H}$. This proves the first part of Proposition 1.

Uniqueness in $\mathcal{H}$ means equality P-almost surely, not a unique coefficient vector or a pointwise extension outside the support. For duplicated peers $Y_j = Y_k$, changing (w_j, w_k) while keeping their sum fixed leaves h_w unchanged. More generally, the set of minimizing weights is the inverse image of one projected response under A. No full-rank assumption on the peer columns is needed for response uniqueness.

The projection satisfies

$$\langle Y_t - h_t^\star, h - h_t^\star \rangle_{L^2(P)} \le 0 \quad \text{for every } h \in \mathcal{C}_t. \tag{7}$$

Consequently,

$$\|Y_t - h\|^2_{L^2(P)} - \|Y_t - h_t^\star\|^2_{L^2(P)} \ge \|h - h_t^\star\|^2_{L^2(P)}. \tag{8}$$

These statements concern squared-loss projection. A convex combination fitted by absolute loss need not induce a unique response.

A.2 Consistency with compatible evaluation support

Let $\|Y_j(z)\|_2 \le B$ for all target and peer responses. For independent observations from P, define

$$L(w) = \mathbb{E}_P \|Y_t - h_w\|_2^2, \qquad \widehat{L}_m(w) = \frac{1}{m} \sum_{i=1}^m \|Y_t(Z_i) - h_w(Z_i)\|_2^2. \tag{9}$$

With finitely many bounded peer responses, expanding the square reduces uniform convergence to convergence of finitely many empirical first and second moments. Thus

$$\epsilon_m := \sup_{w \in \Delta} |\widehat{L}_m(w) - L(w)| \to_p 0. \tag{10}$$

For any empirical minimizer $\widehat{w}$ and population minimizer $w^\star$, empirical optimality gives $L(\widehat{w}) - L(w^\star) \le 2\epsilon_m$. Equation (8) then yields $\|\widehat{h}_t - h_t^\star\|^2_{L^2(P)} \le 2\epsilon_m$.

Let the evaluation sample be independent and distributed according to Q. If $Q \ll P$ and $\mathrm{d}Q/\mathrm{d}P \le \kappa < \infty$, then

$$\mathbb{E}_Q \|\widehat{h}_t - h_t^\star\|_2 \le \sqrt{\kappa}\, \|\widehat{h}_t - h_t^\star\|_{L^2(P)} \to_p 0. \tag{11}$$

The reverse triangle inequality bounds the change in expected residual magnitude by this quantity. Conditional on the fit, evaluation residuals are independent and bounded by $2B$, so their sample mean differs from its Q-expectation by $o_p(1)$. This proves Proposition 2. The same argument tolerates an optimization error that vanishes with sample size. For matched designs with finitely many realizations per question, the same argument applies to question-level averages of the fitting and evaluation losses. Independence is required between question clusters, not between realizations within a cluster.

Support compatibility is essential to this formulation. If peers coincide P-almost surely but differ on a set charged only by Q, their fitted representations can yield distinct Q-residuals. A fixed tie-breaking rule still defines an implementation, but projection uniqueness under P alone does not identify a unique held-out population response. The main endpoint designs put positive mass on both clean and stressed conditions; the conditional evaluation of either is dominated by their mixture design.

A.3 Monotonicity and its limits

For nested peer sets $\mathcal{J}_1 \subseteq \mathcal{J}_2$, assigning zero weights to the added peers embeds the smaller simplex into the larger. Therefore

$$\inf_{h \in \mathcal{C}(\mathcal{J}_2)} \mathbb{E}_P \|Y_t - h\|_2^2 \le \inf_{h \in \mathcal{C}(\mathcal{J}_1)} \mathbb{E}_P \|Y_t - h\|_2^2. \tag{12}$$

This proves the monotonicity assertion in Proposition 1. The same set-inclusion argument applies to any fixed fitting loss optimized over the two classes. It does not compare held-out MAE values of MSE-trained estimators, nor estimators evaluated under different designs. Negative test-set group gains and removal effects are therefore compatible with the framework.

If $\mathcal{C} \subseteq \mathcal{F}$ is a richer admissible response class, its optimized fitting loss cannot be larger. Hence a large convex residual only establishes a gap relative to $\mathcal{C}$, not inexpressibility under arbitrary transformations of peer outputs. Fixed convex combinations and input-dependent single-peer selectors are generally different classes: neither should be identified with the other without further assumptions.

A.4 Aggregation of intervention realizations

For a fixed fit and scalar residuals $R_1, \ldots, R_K$,

$$\left| \frac{1}{K} \sum_{k=1}^{K} R_k \right| \leq \frac{1}{K} \sum_{k=1}^{K} |R_k|. \tag{13}$$

Equality holds when the nonzero residuals have a common sign. Averaging signed responses can therefore conceal within-question disagreement across realizations. The primary estimator averages absolute residuals over intervention tracks. Averaging answer-label rotations is a separate choice: it defines the semantic score that is subsequently audited. Its residual is not the mean residual under each presentation.

B Language-model audit protocol

B.1 Checkpoints and question sample

The ecosystem comprises the eight checkpoints in Table 4. Exact revision identifiers are preserved in the accompanying model manifest. The family sources are Qwen2.5 (Hui et al., 2024), General Reasoner (Ma et al., 2025), Phi-4 and Phi-4 reasoning (Abdin et al., 2024; 2025), Mistral NeMo (Mistral AI, 2024), OLMo 2 (OLMo et al., 2024), Granite 3.3 (IBM Granite Team, 2025), and Gemma 3 (Team et al., 2025). Qwen and Phi are the declared sibling groups; this metadata does not assert behavioral equivalence.

Table 4: Fixed model ecosystem. Short names are used consistently in the figures.

Short name	Checkpoint
Qwen2.5	`Qwen/Qwen2.5-7B-Instruct`
General Reasoner (GR)	`TIGER-Lab/General-Reasoner-Qwen2.5-7B`
phi-4	`microsoft/phi-4`
Phi-R	`microsoft/Phi-4-reasoning-plus`
Mistral	`mistralai/Mistral-Nemo-Instruct-2407`
OLMo	`allenai/OLMo-2-1124-13B-Instruct`
Granite	`ibm-granite/granite-3.3-8b-instruct`
Gemma	`google/gemma-3-12b-it`

The fixed sample contains 560 MMLU-Pro questions, 40 per subject across 14 subjects. Each question retains its original number m_q of valid options; the audit does not assume every row has exactly ten. Ten stratified 50:50 splits contain 280 fitting and 280 evaluation questions each. A question and all of its doses, tracks, model responses, and label rotations remain on the same side within a split. The ten splits reuse the same physical questions and are not independent datasets.

B.2 Matched interventions

Irrelevant-context injection prepends unrelated stems from other categories at nested word budgets $0, 64, 128, 256, 512$. Content deletion replaces selected eligible non-stopword tokens in the question stem with a literal blank at fractions $0, 0.1, 0.2, 0.3, 0.4$. Options and the reference answer are unchanged. Three realizations are constructed per question at each nonzero dose. The semantic intervention is shared across models before model-specific chat wrapping.

B.3 Candidate-label scores and balanced responses

For each valid visible label k, we evaluate the conditional sequence log likelihood $\ell_{j,k}$ of the prescribed short label continuation, without length normalization, and set

$$p_{j,k} = \frac{\exp(\ell_{j,k})}{\sum_{r=1}^{m_q} \exp(\ell_{j,r})}. \tag{14}$$

These scores are conditional on the prompt and answer format. For question q, evaluate all m_q cyclic rotations and remap visible labels to their original semantic options. The balanced response is

$$\bar{p}_j(s) = \frac{1}{m_q} \sum_{r=0}^{m_q-1} p_j^{(r)}(s), \qquad Y_j = \bar{p}_j(y_q^\star). \tag{15}$$

Every semantic option occupies every position exactly once. This balances positions, not all $m_q!$ permutations; it need not remove interaction effects or recover latent beliefs. Because normalization is nonlinear, balancing is not generally equivalent to subtracting an additive logit bias.

B.4 Distinct full-dose and endpoint designs

The canonical-order trajectories fit one vector over all five doses, giving equal mass to each question and dose and dividing nonzero-dose mass equally among the three tracks. In contrast, the balanced-interface comparisons use only clean and maximum-stress conditions. For each family and split, their fitting objective is

$$\widehat{L}_t(w) = \frac{1}{|\mathcal{Q}_{\text{fit}}|} \sum_{q \in \mathcal{Q}_{\text{fit}}} \left[\frac{1}{2} r_q(0; w)^2 + \frac{1}{6} \sum_{k=1}^{3} r_q(d_{\max}, k; w)^2 \right]. \tag{16}$$

Endpoint evaluation uses

$$\widehat{U}_t(0) = \frac{1}{|\mathcal{Q}_{\text{eval}}|} \sum_q |r_q(0)|, \quad \widehat{U}_t(d_{\max}) = \frac{1}{3|\mathcal{Q}_{\text{eval}}|} \sum_q \sum_{k=1}^{3} |r_q(d_{\max}, k)|. \tag{17}$$

Overall endpoint-design errors average the two endpoints with equal mass. Endpoint effects subtract the clean value from the stressed value. The raw endpoint baseline uses this same two-endpoint design before comparing it with balanced responses, avoiding a confound between interface change and omission of intermediate doses.

B.5 Matched fitting losses and peer removal

For the balanced reference-answer analysis, the squared-loss comparison selects the single peer of smallest fitting MSE and compares it with the simplex-constrained MSE fit. The absolute-loss comparison analogously selects by fitting MAE and uses an absolute-loss convex fit. All four errors are evaluated by held-out MAE under the same design. The canonical complete-option diagnostic has its own fitting and evaluation specification in Appendix C.3. A selected single peer is fixed across evaluation inputs. For each deletion of a peer, the reduced model is refitted on the same fitting questions and evaluated on the same held-out questions.

B.6 Common-question bootstrap and selection scope

Each of 1,000 bootstrap replicates draws one subject-stratified multiplicity vector over the 560 base questions. That same vector is restricted to the fitting and evaluation partitions of all ten fixed splits. Each fit, single-peer selection, and relevant peer-removal comparison is recomputed using the weighted sample; split estimates are then averaged. This retains dependence arising when a physical question occurs in several split definitions. Reported percentile intervals are pointwise, not simultaneous family-wise intervals.

The lineage comparisons and focused validation targets were fixed before the balanced-interface evaluation, following inspection of the earlier canonical audit. The new interface measurements use the same question pool, not a fresh independent benchmark. The six-of-eight summary is an ecosystem-wide descriptive result; it is not presented as six independently preregistered discoveries.

B.7 Generation as an auxiliary interface

A fixed 140-question subset, ten per subject, is evaluated by deterministic generation with a 512-token budget under clean and two maximum-dose conditions. All cyclic rotations are included and extracted answer labels are remapped to semantic options. These measurements are not inputs to the primary PIER estimator.

The archived parser accepts explicit final-answer patterns but also has a permissive fallback for a valid standalone label near the end of an output. In long reasoning sequences, this can identify an intermediate label as an answer. Hitting the token budget also need not imply successful completion. Accordingly, the archived agreement and malformed rates in Appendix C are diagnostics under that parser, not verified final-answer accuracy or proof that truncation was resolved.

C Complete language-model comparisons and diagnostics

C.1 Balanced endpoint effects

Table 5 includes every target and both intervention families, rather than only the headline directions. A positive effect means that the fixed fitted peer combination has greater residual at maximum stress. Agreement of signs across overlapping splits is descriptive stability, not a binomial test with ten independent repetitions.

Table 5: Balanced endpoint changes with pointwise 95% common-question bootstrap intervals.

Target	Irrelevant context	Content deletion
Qwen2.5	$+0.0144$ $[+0.0073, +0.0209]$	$+0.0064$ $[-0.0011, +0.0142]$
GR	$+0.0030$ $[-0.0032, +0.0091]$	$+0.0031$ $[-0.0039, +0.0097]$
OLMo	-0.0103 $[-0.0171, -0.0031]$	-0.0157 $[-0.0257, -0.0062]$
Gemma	$+0.0108$ $[+0.0018, +0.0195]$	$+0.0004$ $[-0.0125, +0.0119]$
Granite	-0.0050 $[-0.0132, +0.0024]$	-0.0056 $[-0.0155, +0.0037]$
Phi-R	$+0.0257$ $[+0.0161, +0.0353]$	$+0.0038$ $[-0.0043, +0.0109]$
phi-4	$+0.0325$ $[+0.0171, +0.0444]$	$+0.0034$ $[-0.0107, +0.0153]$
Mistral	-0.0089 $[-0.0165, -0.0018]$	-0.0218 $[-0.0319, -0.0127]$

Gemma and Phi-R have positive irrelevant-context effects whose intervals exclude zero, but these are not treated as generated-behavior conclusions. Granite's balanced effects cross zero despite stable negative split signs. OLMo's deletion effect remains a scalar-interface result; it should not inherit stronger full-output claims from the Mistral result.

C.2 Single-peer versus collective coverage

Table 6 gives the relative held-out improvement for every target and each matched fitting loss. These percentages are computed from the mean single and group errors; uncertainty is reported for the underlying absolute improvement. Intervals for both losses and complete removal results are supplied as machine-readable tables.

Table 6: Relative held-out group gains (percent). D: deletion; I: irrelevant context. Absolute-improvement intervals are preserved in the accompanying CSV.

Target	D / MSE	D / MAE	I / MSE	I / MAE
Qwen2.5	-0.87	0.08	-0.82	-0.20
GR	15.30	16.20	17.96	18.34
OLMo	4.68	6.06	5.85	7.89
Gemma	10.11	12.42	16.46	14.87
Granite	9.75	10.74	12.03	12.45
Phi-R	31.53	29.65	25.08	23.80
phi-4	7.46	8.94	12.56	10.94
Mistral	19.40	20.43	18.82	18.79

The Qwen pair's asymmetry does not concern its direct pairwise error, which is symmetric under the matched absolute response difference. It concerns the improvement supplied by the remaining peers

Table 7: Declared sibling-removal effects. The comparison column is the largest mean effect of deleting one non-sibling.

Target	Family	Sibling effect [95% CI]	Non-sibling max
Qwen2.5	D	$0.0479\ [0.0398, 0.0583]$	0.0006
Qwen2.5	I	$0.0450\ [0.0354, 0.0567]$	0.0008
GR	D	$0.0406\ [0.0333, 0.0490]$	0.0020
GR	I	$0.0401\ [0.0321, 0.0496]$	0.0026
Phi-R	D	$0.0112\ [0.0071, 0.0161]$	0.0058
Phi-R	I	$0.0083\ [0.0041, 0.0124]$	0.0069
phi-4	D	$0.0082\ [0.0026, 0.0140]$	0.0055
phi-4	I	$-0.0016\ [-0.0058, 0.0034]$	0.0081

when either member is the target. Small and sometimes negative Phi sibling-removal effects provide a contrast to the strong Qwen dependence.

C.3 Complete-option-distribution audit

The complete-option analysis uses the same 560 questions and ten stratified half-splits as the canonical scalar audit. It retains the canonical option order and all five doses in each family: $0, 64, 128, 256, 512$ words for irrelevant context and $0, 0.1, 0.2, 0.3, 0.4$ for deletion. Each question–dose pair has equal design mass; the three nonzero-dose tracks divide that mass equally. The response contains the probabilities of every valid option, with zero padding in the fitting arrays where option counts differ.

Two weight sources are evaluated. *Transferred* weights are learned from scalar reference-answer responses and applied to the entire vector. *Refitted* weights minimize the weighted sum of squared vector residuals over the fitting questions. A single peer is selected by its smallest mean fitting total variation. All fits are fixed over evaluation inputs. This is a squared-vector-fit/TV-evaluation diagnostic; its single-selection objective differs from group fitting, unlike the loss-matched scalar comparisons in the main experiment.

For valid option distributions p and q, the reported distance is

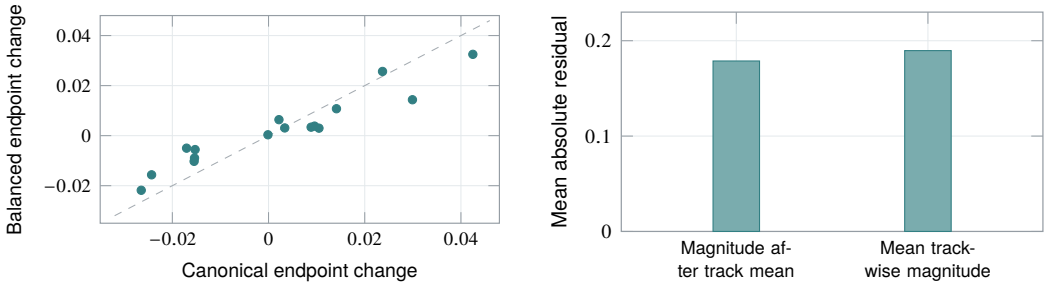
$$\mathrm{TV}(p, q) = \frac{1}{2} \sum_{s} |p(s) - q(s)|. \tag{18}$$

The archived implementation clips valid coordinates to $[10^{-15}, 1]$ and renormalizes before evaluating TV, Jensen–Shannon divergence, semantic top-1 agreement, and correctness agreement. It averages within each dose and then across the five equally weighted doses. The supplied table contains 1,600 rows: eight targets, two families, ten splits, two weight sources, and five doses. Overall values repeated on the five dose rows are counted only once when summarizing splits.

The single peer selected for each Qwen target is its sibling in all ten splits and both families. This produces equal pairwise TV in both directions, while the effect of the additional peers differs. For General Reasoner, vector refitting improves mean TV by 4.9% under deletion and 9.6% under irrelevant context. For Qwen2.5, the group's TV is higher. Mistral improves by 14.3% and 15.4%, and Phi-R by 18.0% and 16.2%. Gemma's vector group gains have the opposite sign from its balanced scalar results. Table 8 includes these differing cases together with all other targets.

The vector sibling-removal diagnostic refits the remaining peers by the same vector squared loss. Both Qwen directions show positive TV inflation in every split (Table 3); the Phi directions have smaller and less uniform effects. This diagnostic does not include a complete vector non-sibling removal null. Likewise, the tables report split means and direction counts, not new bootstrap intervals. The overlapping splits are a description of stability within the fixed question pool, rather than ten independent replications.

The canonical vector design and the balanced scalar endpoint design differ in response representation, presentation, dose support, and fitting comparisons. Their common patterns establish evidence under both recorded views; differences do not isolate any one of those choices. The balanced-interface common-question intervals are reported only for the measurements to which they apply.



(a) Endpoint changes under two interfaces. **(b)** Order of aggregation under a fixed fit.

Figure 5: Representation and aggregation are separate audit choices. Left: all 16 endpoint changes under matched endpoint designs. Right: mean absolute residual after versus before averaging intervention tracks, with the trackwise-fit weights held fixed. Neither comparison adds new model observations.

C.4 Interface rankings and intervention aggregation

Rankings of the eight targets' family-averaged endpoint-design MSE-fit errors have canonical-versus-balanced Spearman correlation 0.952. Across the 16 target–family endpoint changes, the corresponding correlation is 0.932, and 15 signs agree. These statistics refer to different objects and should not be interchanged.

For a fixed fit, measuring each realized intervention and then averaging gives $K^{-1}\sum_k |R_k|$, whereas first averaging the signed responses gives $|K^{-1}\sum_k R_k|$. Jensen's inequality orders these quantities; equality holds when the nonzero residuals share a sign. The first retains variation across matched intervention realizations, while the second measures reconstruction of an averaged response.

In the canonical audit, fitting and evaluating aggregate responses gives mean residual 0.17849, compared with 0.18964 under trackwise fitting and evaluation. This 6.25% difference includes refitting. Figure 5(b) isolates the evaluation order using identical trackwise-fit weights. These calculations are diagnostic comparisons of estimands, separate from the choice to define a label-balanced response before fitting.

C.5 Archived generation readouts

The table averages stored per-condition readouts. Score–generation agreement is measured across individual rotations, not between one balanced argmax and every generated answer. The permissive

Table 8: Complete-option reconstruction in the canonical audit. Each cell averages ten fixed splits. D/I denote deletion/irrelevant context. Single denotes selection by fitting TV; transferred uses scalar-fit weights to predict every option; refitted minimizes vector squared error. + is the number of splits with single TV greater than refitted-group TV, not an independent-replication test.

Target	Family	Single TV	Transferred TV	Refitted TV	Gain (%)	+
Qwen2.5	D	0.3140	0.3387	0.3423	-9.0	0/10
Qwen2.5	I	0.3372	0.3629	0.3630	-7.6	0/10
General Reasoner	D	0.3140	0.2935	0.2986	+4.9	10/10
General Reasoner	I	0.3372	0.3009	0.3048	+9.6	10/10
OLMo	D	0.4706	0.4500	0.4479	+4.8	10/10
OLMo	I	0.4573	0.4381	0.4360	+4.7	10/10
Gemma	D	0.4856	0.5402	0.5429	-11.8	0/10
Gemma	I	0.4682	0.5443	0.5383	-15.0	0/10
Granite	D	0.5028	0.4740	0.4745	+5.6	10/10
Granite	I	0.4845	0.4627	0.4653	+4.0	10/10
Phi-R	D	0.4205	0.3472	0.3447	+18.0	10/10
Phi-R	I	0.4164	0.3536	0.3490	+16.2	10/10
phi-4	D	0.4612	0.4331	0.4385	+4.9	10/10
phi-4	I	0.4595	0.4495	0.4585	+0.2	4/10
Mistral	D	0.4205	0.3630	0.3604	+14.3	10/10
Mistral	I	0.4164	0.3557	0.3523	+15.4	10/10

Table 9: Archived generation diagnostics, averaged across the three conditions. Values depend on the permissive answer parser and are not verified final-answer accuracy.

Model	Per-rotation agreement	Malformed fraction
Qwen2.5	0.846	0.001
GR	0.334	0.012
OLMo	0.750	0.000
Gemma	0.926	0.007
Granite	0.569	0.002
Phi-R	0.275	0.001
phi-4	0.445	0.007
Mistral	0.647	0.002

parsing fallback and frequent budget-length outputs require particular caution for reasoning models. We do not use malformed rates as evidence of reliable final-answer completion.

C.6 Numerical feasibility and environment

The balanced-interface absolute-loss bootstrap uses a deterministic post-solve simplex projection for otherwise accepted solutions with raw violations no greater than 10^{-7}. The repair was explicitly authorized after a feasibility-gate failure. Across 394 repaired solutions, the largest objective change was 1.46×10^{-9} and the largest held-out prediction change was 1.01×10^{-7}. A repair-free-only diagnostic did not change the recorded headline decisions; it is a sensitivity check, not an alternative primary bootstrap.

Post-processing used CPython 3.11.15 instead of the prestaged 3.11.13, with the locked package versions preserved. This is not a bitwise-identical runtime. The original inference outputs, split definitions, and bootstrap seeds were retained. The accompanying provenance record distinguishes this numerical implementation adjustment from any change of the scientific response definition.

D Vision and forecasting protocols

D.1 Separate vision ecosystems and score interfaces

The adversarial audit contains ResNet-18, ResNet-50, EfficientNet-B0, ConvNeXt-Tiny, ViT-B/16, and a robust ResNet-50. It uses 500 images with an even fitting/evaluation split. The source implementation applies a common ResNet-50-referenced gradient-sign perturbation at doses $0, 0.02, 0.04, 0.06, 0.08, 0.10$ and records a bounded margin score. These dose units follow the archived normalized-input implementation, not a separately verified pixel-space robustness budget.

The seven-model context audit contains ResNet-50, EfficientNet-B0, ConvNeXt-Tiny, ViT-B/16, and three shape-trained ResNets. The variants are trained on stylized ImageNet (SIN), SIN and ImageNet (SIN+IN), or SIN+IN followed by ImageNet fine-tuning (SIN+IN$\rightarrow$IN). Each context has 800 images split into 400 fitting and 400 evaluation examples. This audit records the true-class probability and fits separately in each context. Its numerical residual scale is not comparable with the adversarial margin audit.

The separate five-model variant experiments retain the same four standard peers plus one shape-trained variant. Their repeated standard-target rows describe different peer sets, not independent replicates to average silently. The main context panel uses the seven-model table only. The PCA experiment uses a shape-trained target with the four standard peers and the fitting response vectors.

D.2 Projected response geometry

For each target and context, the archived PCA input consists of four peer vectors, one target vector, and the fitted mixture vector, each evaluated on 400 fitting images. The first two principal components are fitted separately for each target–context combination. Figure 6 displays all six archived projections; they cannot be interpreted as points in one shared coordinate system.

A common affine PCA map preserves convex combinations with coefficients summing to one. Exclusion from the exact projected peer hull therefore implies exclusion from the original peer hull;

projected inclusion does not establish original-space inclusion. The stored mixture is the projection of the original-space fitted mixture, not a fresh nearest-point solution in two dimensions. Its displayed segment need not be the shortest segment to the two-dimensional hull. Archived numerical weights have small feasibility deviations, which are retained rather than silently repaired.

Table 10: Archived fitting geometry and support coefficients. Coefficients are rounded only for display; exact values and feasibility residuals remain in the data.

Target	Context	RN-50	EffNet	ConvNeXt	ViT	L_2 residual
SIN	Natural	-0.000	-0.000	-0.000	1.000	0.00488
SIN	Shape	0.507	0.223	0.000	0.270	0.04673
SIN+IN	Natural	0.000	1.000	0.000	0.000	0.01076
SIN+IN	Shape	0.150	0.139	0.000	0.711	0.02076
SIN+IN→IN	Natural	-0.000	-0.000	-0.000	1.000	0.00604
SIN+IN→IN	Shape	-0.000	0.000	0.813	0.187	0.03309

D.3 Residual concentration and matched ordering

For non-negative residual magnitudes $r_1, \ldots, r_n$, sort $r_{(1)} \geq \cdots \geq r_{(n)}$ and compute cumulative mass $\sum_{i=1}^{k} r_{(i)} / \sum_{i=1}^{n} r_i$. The main curves include every integer k from zero to 400. Top-k overlap is the size of the intersection of two highest-residual index sets divided by k. Sorting ties uses a fixed stable index order.

The shared pair is specifically ConvNeXt-Tiny and SIN+IN→IN, whose archived identifier is `ShapeResNet50_ShapeResNet`; it is not the SIN target illustrated in the main PCA panels. The seven-model arrays yield top-20 overlap of 19 under shape-biased evaluation and two under natural evaluation. The absolute-residual correlation is 0.9947 in the former setting. These arrays contain residuals and intervention levels but no image identifiers. The source code builds the evaluation ordering before the per-target loop; this supports indexwise alignment, while limiting independent file-level identity verification.

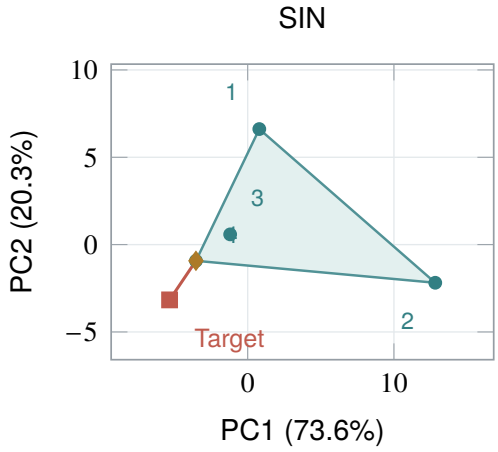
D.4 Forecasting comparisons and numerical scope

The 31-city summary records locally trained forecasting models, a pooled global model, and a peer fit for each target. It reports 4,000 fitting and 4,000 residual-evaluation windows per city. Stored task-error summaries and residual summaries need not cover identical numbers of windows, so they are compared at city level rather than treated as paired per-window estimates.

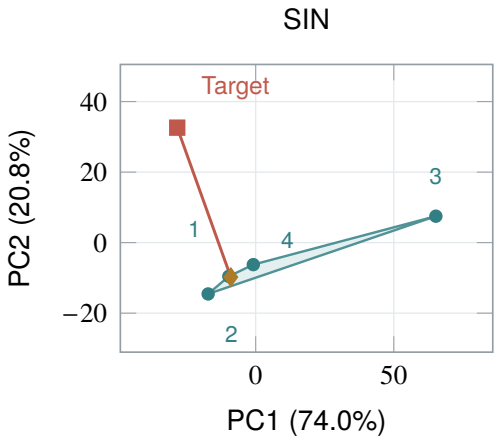
The normalized axes in Figure 7 are $100\,\mathrm{PIER}/\mathrm{MeanFlow_Test}$ and $100\,\mathrm{Delta_Router}/\mathrm{Local_MAE}$. Their correlation is 0.05188; the correlation of raw PIER and raw task-loss change is 0.83597. The stored group error is lower than the closest single-peer diagnostic for 26 cities and higher for five. The closest-peer field minimizes recorded response discrepancy and is not documented in these artifacts as a fitting-only selection. It must not be equated with the honest matched-objective single-peer baseline used in the LLM study.

The long-format weight file has 961 rows. Its weights are numerically approximate: the minimum is -0.004825 for Luzern and -0.004698 for Paris; Luzern's stored top-three mass is greater than one. These are material departures from an exact simplex constraint, not rounding to a probability vector. The raw outputs are preserved, and no strict-convex certificate or support-probability interpretation is claimed for this archive. Original response matrices needed to recompute predictions after repair are not supplied here.

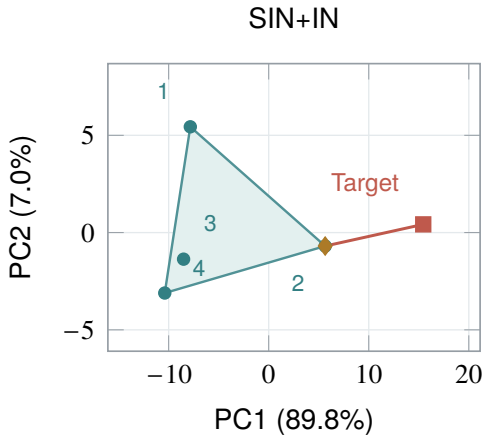
The archive supports static city-level reconstruction and task-error comparisons. Sequential removal would change the available peer sets and require corresponding refits and predictions; no such trajectory is reported here.



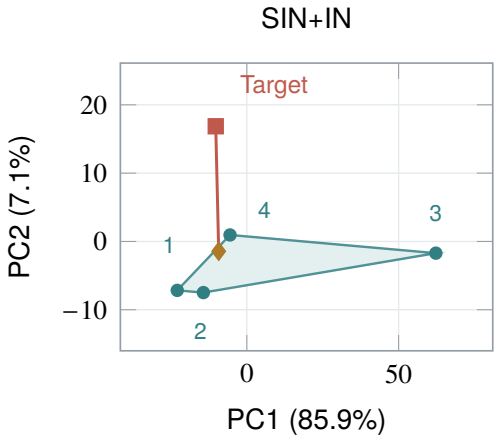
(a) SIN / natural.



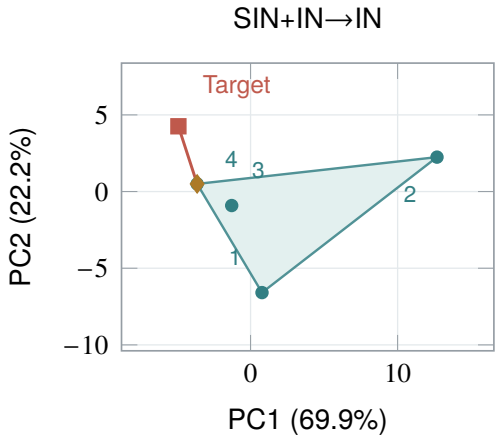
(b) SIN / shape-biased.



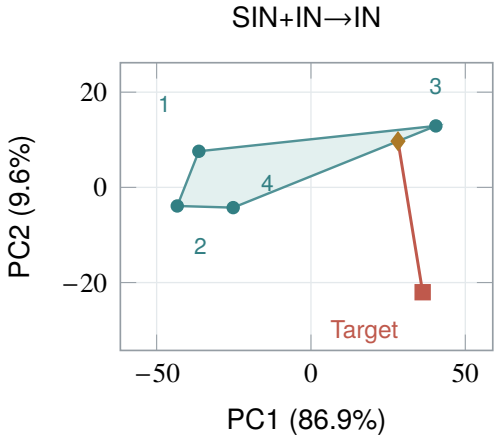
(c) SIN+IN / natural.



(d) SIN+IN / shape-biased.

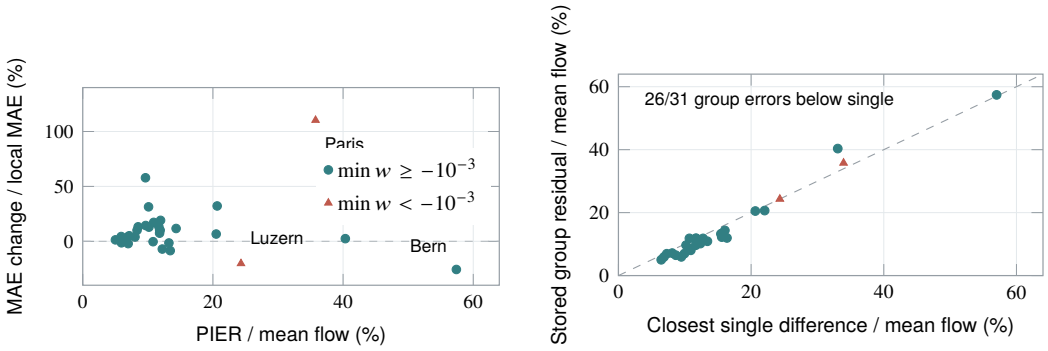


(e) SIN+IN→IN / natural.



(f) SIN+IN→IN / shape-biased.

Figure 6: All archived PCA views, fitted separately. Peer labels: 1 ResNet-50; 2 EfficientNet-B0; 3 ConvNeXt-T; 4 ViT-B/16. Gold diamonds denote the archived fitted mixture. Coordinates in each panel are multiplied by 10^3 for readability; bases differ, so displacement across panels is not geometrically comparable.



(a) Relative residual and task-loss change. **(b)** Stored peer fit versus closest single response.

Figure 7: Reconstruction and task utility are different comparisons. Each point is one city in the archived table. Both axes in (b) are normalized by that city's mean flow; the dashed line denotes equal error. The stored group fit has lower error in 26 of 31 cities, not all 31. Some stored coefficients violate the simplex constraint; the points describe measured fits, not exact convex certificates (Appendix D.4).

E Additional distinctions and query bounds

E.1 Unmatched observations and reference support

Suppose logs for all models are supported on S_0, while the reference audit distribution $P_\star$ assigns positive mass to $S \setminus S_0$. Consider peers that are zero everywhere. In one ecosystem the target is also zero; in a second it equals one on $S \setminus S_0$ and zero on S_0. The entire observational law is identical, but reference-design PIER is zero in the first ecosystem and positive in the second. Thus logs lacking reference support do not identify the audit residual in general. This construction does not rule out identification from observational data with sufficient shared support and additional assumptions.

E.2 Utility credit is not coverage redundancy

Consider two identical model responses and a coalition game worth one whenever at least one model is available, and zero otherwise. Each model has Shapley value $1/2$, while either is exactly covered when the other is its peer. Positive utility credit can therefore coexist with zero PIER. This counterexample distinguishes the objects measured; it is not a criticism of Shapley values for their intended utility-allocation task.

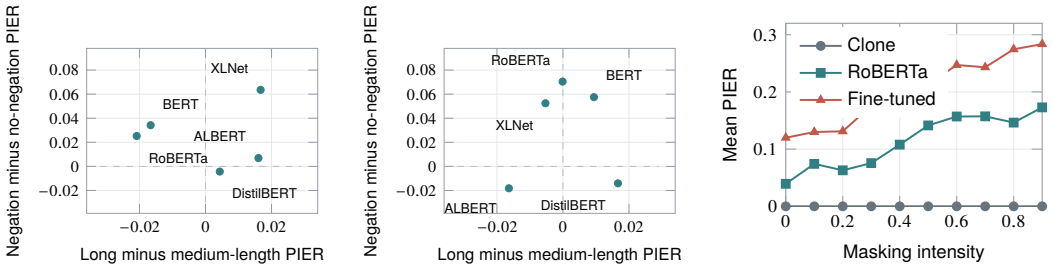
E.3 A qualified finite-design query bound

Assume a local linear response model $Y_j(z) = \phi(z)^\top \beta_j$ with d features. Choose d audit points with invertible feature matrix Φ. In the noiseless setting, one response at each point recovers β_j for every model. The coefficient representation then reduces exact coverage to membership in the convex hull of peer coefficients, provided the features identify responses on the reference design.

Under independent sub-Gaussian noise with variance proxy σ^2, averaging r repetitions at each design point controls response-mean errors by a union bound. Multiplication by Φ^{-1} transfers this error to the coefficients. For a fixed conditioning-dependent constant C_Φ, a sufficient repeat count to resolve a separated margin γ scales as

$$r \geq C_\Phi \frac{\sigma^2}{\gamma^2} \log\left(\frac{Nd}{\delta}\right). \tag{19}$$

The query count is dr per model, or Ndr for the complete ecosystem. In the one-dimensional Gaussian case, testing mean zero against a mean separated by γ requires $\Omega(\sigma^2\gamma^{-2}\log(1/\delta))$ observations. This matches the dependence on noise, separation, and confidence in that worst case, not a general lower bound in d or N. No full-dimensional minimax-optimality claim is made.



(a) Subgroup contrasts at dose zero. **(b)** Subgroup contrasts at dose 0.5. **(c)** Constructed target responses.

Figure 8: Subgroup localization and exact-duplication checks. Fingerprint coordinates are arithmetic differences of archived context means; lines in (c) connect the observed dose-grid values. No missing confidence interval is reconstructed. The clone residual is retained at its actual nonzero numerical value, although it is visually indistinguishable from zero on the common scale.

E.4 Residual magnitude does not identify robustness

Let the intervention level be uniform on $[0, 1]$. A constant residual $R_0(\theta) = c$ and the family $R_m(\theta) = (\pi c/2)\sin(2\pi m\theta)$ all have mean absolute residual c. Their derivatives, however, have magnitudes growing with m. Average PIER therefore does not determine sensitivity along the intervention path. This motivates reporting trajectories and input-level concentration in addition to an average gap.

F Additional sentiment-classifier audits

The SST-2 experiments complement the main generative-model study by examining identifiable input subgroups and constructed target relations. They contain BERT, DistilBERT, RoBERTa, ALBERT, and XLNet classifiers under matched token masking. They are not described as a modern generative-model benchmark.

Context contrasts. For each model and dose, the horizontal fingerprint coordinate subtracts mean PIER on the reported medium-length bin (8–15) from that on the long-sentence bin (more than 15); the vertical coordinate subtracts the no-negation group from the explicit-negation group. Figure 8(a)–(b) uses the stored subgroup means at doses 0 and 0.5. ALBERT moves from positive contrasts at dose zero to negative contrasts at dose 0.5. DistilBERT has a negative negation contrast at both doses but a positive length contrast. These are descriptive subgroup contrasts, not causal effects of sentence length or negation.

Constructed targets. A separate fixed-peer audit compares a DistilBERT clone, RoBERTa, and a fine-tuned DistilBERT. The clone's maximum residual is 1.31×10^{-8} across the masking grid. RoBERTa and the fine-tuned target have nonzero gaps that generally rise with dose (Figure 8c). The fine-tuned target already has a residual of 0.120 at baseline; stronger masking increases an already present gap. These controls establish different response relations without attributing the observed differences solely to architecture or a particular training mechanism.