A Benchmarked Diagnostic for Sparse Decomposability of Dense Causal Subspaces

Socrates Osorio [*1] **Joy Zheyun Yang** [*2 3]

Abstract

Benchmarks for mechanistic interpretability should test not only whether a causal variable can be localized, but whether the localized subspace can be recovered in a reusable representation basis. We propose sparse decomposability as a benchmarked diagnostic for dense causal subspaces: given a DAS-style teacher and a fixed pretrained SAE dictionary, causal sparse distillation (CSD) measures how much interchange-intervention behavior survives when the intervention is constrained to a small set of SAE latents. The diagnostic is calibrated on a 16-cell synthetic benchmark with ground-truth supports, where CSD-L1 recovers correlated-distractor support ($\mathrm{F1} = 1.00$) while DBM and DiffMean controls fail. On dense-valid Gemma/Qwen tuples, two compact pre-CSD decoder-geometry statistics predict CSD/dense recovery with leave-one-out $R^2 = 0.89$ and bootstrap 95% CI $[0.79, 0.95]$, while model size alone gives $R^2 = -2.00$. Public-harness evaluations and matched diagnostic controls then separate positive selector-specific MCQA cases from random-K-degenerate RAVEL rows and sparse-limited 27B sites. Matched random-K controls show that high CSD/dense recovery is not by itself evidence of meaningful feature selection. The result is a benchmark-facing instrument: it maps where dense causal variables are sparse-decomposable in existing SAE bases and where benchmark scores should not be interpreted as SAE-level explanations.

[*]Equal contribution [1]Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA, USA [2]University of Oxford, Oxford, United Kingdom [3]Department of Computer Science, Georgetown University, Washington, DC, USA. Correspondence to: Socrates Osorio <socratesj.osorio@berkeley.edu>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

1. Introduction

Mechanistic interpretability has two useful but mismatched tool families. Supervised causal localizers, especially DAS (Geiger et al., 2024), optimize directly for interchange intervention accuracy (IIA) and can identify faithful subspaces. Their outputs, however, are dense learned directions. SAEs provide a dictionary of sparse, often human-labeled latents, but MIB reports that SAE features are not generally competitive with DAS on causal variable localization (Mueller et al., 2025). This mismatch matters because the field often treats SAE failures and DAS successes as evidence about different objects: one about learned dictionaries, the other about causal subspaces. If sparse decomposability itself changes with scale or intervention site, then neither conclusion is stable without measuring whether the dense causal signal is representable in the available SAE basis.

This paper asks a measurement question rather than assuming either tool is always preferable: how much of a faithful DAS intervention can be expressed as a small set of pretrained SAE latents, and when does that expression fail? The answer is tuple-specific. Our strongest behavioral-recovery result is on RAVEL continent interventions: an SAE-coordinate patch matches dense DAS in a five-seed Gemma-2-9B check using the same public MIB harness, but matched random-K controls make it a coordinate-recoverability result rather than selector-specific feature discovery. Gemma-2-2B ranges from a low aggregate ratio of 0.63 to an independent seed-4 ratio of 1.002. Pilot Gemma-2-27B runs first exposed a weak dense ceiling, and a targeted dense-rescue run then found a valid layer-22 entity-token dense site (mean dense IIA 0.667), but CSD recovers only 42% of that dense intervention. This separates dense-site validity from sparse decomposability at 27B. We introduce causal sparse distillation, a teacher-student procedure that selects SAE features by matching the paired counterfactual intervention behavior of a DAS-style teacher. The output is a faithfulness-sparsity curve and a set of candidate features for audit.

The novelty is the operational test, not the existence of another sparse patch. CSD treats a pretrained SAE basis as fixed and asks whether a dense causal subspace can be compressed into that basis without losing interchange inter-

vention behavior. A high score means the dense variable is sparse-decomposable in the available dictionary. A low score is evidence about the mismatch between that dictionary and the causal variable, not merely an optimization failure.

We use *tuple* to mean a fixed model, layer, token/site, SAE dictionary, causal task, and dense teacher. CSD scores should not be compared across tuples unless the dense-validity gate and sparse baselines are matched.

Core claim. CSD turns sparse decomposability into a benchmarked quantity: a dense causal variable either retains its intervention behavior in a fixed SAE basis, or the benchmark exposes the mismatch.

CSD instantiates a theory–benchmark loop: sparse decomposability is the theoretical property, synthetic cells test ground-truth support recovery, public MIB harnesses test realistic intervention behavior, and decoder geometry gives an ex ante predictor of when the benchmark should succeed.

The remaining evidence maps the boundary of this diagnostic result. MCQA answer-pointer behavior is recoverable with a compact SAE subset on the canonical split but remains stress-fragile under answer-format shifts. A small Llama-3.1-8B check shows why CSD must be read as a diagnostic: answer-symbol interventions are invalid on the canonical MCQA split, but a full-vector sweep identifies the last prompt token as a valid site where CSD again matches dense DAS. IOI is mostly carried by SAE reconstruction-error movement at the tested site, so named SAE latents alone are not enough. Arithmetic carry remains a negative-control case: under our binary first-answer-token proxy, even full-vector and teacher interventions are weak. The held-out arithmetic test result confirms this failure mode: best weighted CSD reaches only 0.122 IIA over 98 usable examples, while full-vector patching is 0.143.

Table 1 summarizes the main claims and their evidential status, including claims the experiments do not support. This explicit accounting is central to the contribution: CSD is a diagnostic for sparse decomposability, not a guarantee that every dense causal subspace has a clean sparse explanation.

What we do not claim. We do not claim monotonic scaling, leaderboard state of the art, uniformly semantic features, or that CSD always beats random-K. We claim that CSD is a diagnostic for when dense causal behavior is recoverable in a fixed SAE basis, with matched null controls determining whether recovery is selector-specific.

2. Related Work

MIB defines the causal variable localization track and measures IIA on counterfactual interventions (Mueller et al., 2025). Its central empirical tension motivates this work: DAS is strong but opaque, while SAE-feature baselines are inspectable but usually weaker. Gemma Scope provides the frozen pretrained SAE dictionary used here (Lieberum et al., 2024). Sparse Feature Circuits rank SAE features by attribution-style influence on model behavior (Marks et al., 2025). Our selection signal is different because it comes from paired interchange interventions generated by a teacher subspace.

SAE motivation builds on superposition and dictionary-learning accounts of polysemantic representations (Elhage et al., 2022; Bricken et al., 2023; Cunningham et al., 2024; Templeton et al., 2024). We also build on causal-intervention circuit work, including IOI, ACDC, and causal scrubbing (Wang et al., 2023; Conmy et al., 2023; Chan et al., 2022). Feature labels and top-activation exemplars are used only as audit evidence, following automatic neuron-description and Neuronpedia-style workflows (Bills et al., 2023; Neuronpedia, 2023).

Closest in spirit are supervised feature-selection methods that improve SAE utility by adding task information, including output-effect steering selection (Arad et al., 2025) and benchmark results showing that simple steering baselines can beat raw SAE features (Wu et al., 2025). Those methods usually optimize scalar output effects or attribution-style rankings. CSD instead optimizes paired counterfactual agreement with a dense causal teacher and reports how much of the teacher survives at each sparsity level. The contribution is therefore not a new ranking heuristic alone, but an operational test of whether an opaque causal subspace is expressible in a fixed, named SAE basis.

3. Method

Each training item consists of a base prompt, a source prompt, and the target counterfactual label. At layer ℓ and token position p, we cache base and source residual-stream activations h_b, h_s, SAE latents z_b, z_s, and the base reconstruction residual

$$e_b = h_b - D_{\mathrm{SAE}}(E_{\mathrm{SAE}}(h_b)).$$

For a relaxed feature mask $s \in [0,1]^M$ over the SAE dictionary, CSD patches selected source latents into the base latent vector:

$$z' = (1-s) \odot z_b + s \odot z_s, \qquad h' = D_{\mathrm{SAE}}(z') + e_b.$$

The residual carryover is essential: without it, evaluation confounds feature selection with SAE reconstruction loss.

The primary training objective combines behavior and

Table 1. Claims and evidence. Supported means the current experiments directly support the claim. Qualified means the evidence supports the diagnostic workflow but should not be read as a standalone headline. Limitation means the result is not claimed as solved.

Claim	Evidence	Status
Tuple-specific SAE-coordinate behavioral recovery	RAVEL CSD/dense is 1.000 across five Gemma-2-9B seeds, but matched random-K controls are competitive, so this supports coordinate-space recoverability rather than selector-specific feature discovery	Qualified
Diagnostic failure modes	Llama stress-matched random-letter gives mean CSD 0.904 vs. dense 0.93 across five seeds and two layers	Supported
Synthetic calibration	Known-support battery gives CSD min positive support F1 1.00 and dense/out-of-span false-positive rate 0.00	Supported
Site validity matters	Llama MCQA succeeds only after full-vector patching identifies a valid token site	Supported
Semantic audit	displayed audit rows give 4/10 majority-plausible features, so causal recovery is not automatically semantic cleanliness	Mixed
Beyond 9B monotonicity	27B dense rescue finds a valid layer-22 entity-token dense site (mean dense 0.667), but CSD recovers only 0.28 IIA / 0.42 of dense	Limitation

teacher-subspace matching:

$$\begin{aligned}\mathcal{L} = &\ \mathrm{CE}(M(\mathrm{do}(h_{\ell,p} \leftarrow h')), y_{\mathrm{cf}}) \\ &+ \lambda_{\mathrm{emb}} \left\| P_T(h' - h_b) - P_T(h_s - h_b) \right\|_2^2 \\ &+ \lambda_1 \|s\|_1,\end{aligned}$$

where P_T is the teacher subspace projection. After optimization, we harden to top-K features and optionally refit per-feature patch weights. We also report CSD-OMP, a deterministic regression-style variant. We report full-vector patching, DBM-SAE, random-K, activation ranking, DiffMean-to-SAE, Arad-style output influence, and SFC-style gradient-activation ranking where available in the run outputs.

4. Experimental Setup

We target Gemma-2-2B with Gemma Scope residual SAEs, using the canonical 16k-width residual release. The cross-family appendices use Qwen-Scope residual SAEs (Qwen Team, 2026) and EleutherAI Llama-3.1-8B residual SAEs (EleutherAI, 2024) where noted. The main MIB-style tasks are MCQA answer pointer, IOI, and RAVEL continent. Arithmetic carry is used as a challenge/negative control. The cross-scale check repeats MCQA and RAVEL on Gemma-2-9B with matched Gemma Scope residual SAEs at the same task sites.

The reported runs are validation and public-test results from the public data path. Section 6 reports the public-test evaluator and aggregator scripts used by the MIB causal-variable track, run against the packaged CSD artifacts.

All site-level results follow the same decision rule. Full-vector patching first tests whether the chosen layer and position can carry the target causal variable at all. Dense DAS then tests whether an optimized dense subspace can control that variable at the valid site. CSD is evaluated only as the next question: how much of that dense causal signal survives when expressed through a small set of fixed pretrained SAE latents?

5. Results

The validation curve in Figure 1 establishes the first sparse-decomposability boundary. MCQA is the positive case: a compact weighted CSD patch reaches full-vector-level IIA while random and DBM-SAE controls remain lower. RAVEL is only partially sparse-decomposable at 2B. IOI identifies a different failure mode: behavior is recoverable when the SAE reconstruction error moves, but not from named SAE latents alone. Arithmetic remains a negative control under the current binary first-answer-token proxy. Tables 13–22 give the detailed controls, seed stability, site ablations, and qualitative intervention examples.

The decoder-geometry claim uses 27 dense-valid Gemma/Qwen model–site–dictionary tuples. Before running CSD, we compute compact predictors from the frozen SAE decoder and the dense teacher: top-K decoder-subspace alignment, stable-rank-style concentration, and model size. Linear predictors are evaluated by leave-one-out cross-validation. Table 2 shows that compact geometry predicts sparse recovery while model size alone does not.

Benchmark validity threats. The benchmark should be read through five validity gates: matched random-K degeneracy, weak dense ceilings, site invalidity, stress-split mismatch, and semantic-label ambiguity. These gates are not post-hoc excuses. They determine whether a high score supports selector-specific sparse recovery, merely SAE-coordinate behavioral recovery, or no reliable sparse-decomposability claim.

Synthetic calibration. The synthetic battery is the ground-truth half of the benchmark loop. It tests planted sparse teachers, correlated distractors, dense-rotated controls, and

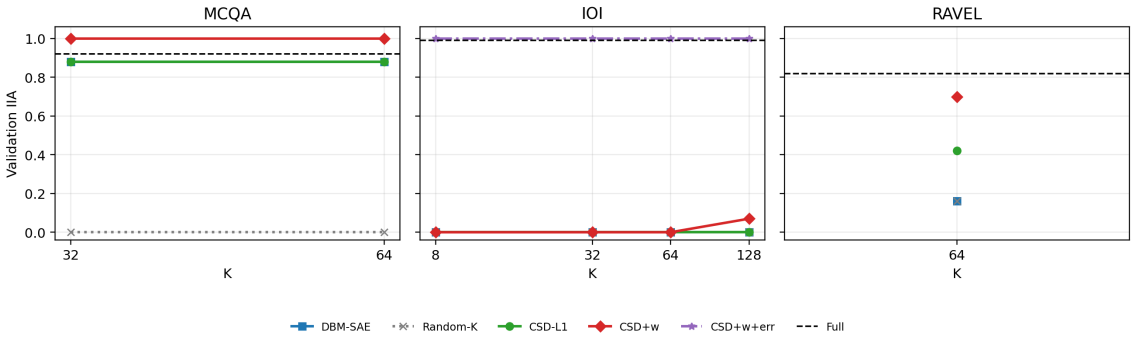

Figure 1. Validation IIA as a function of selected SAE feature count. MCQA shows selector-specific sparse recovery. RAVEL shows SAE-coordinate recovery but requires random-K controls to distinguish true feature selection from redundant-site degeneracy. IOI only recovers when the source-error delta is explicitly allowed.

Table 2. Pre-CSD dictionary-geometry predictor ablation over 27 dense-valid Gemma/Qwen tuples. Compact geometry combines decoder-subspace alignment and stable-rank concentration. Model size alone has negative leave-one-out performance.

Model	N	LOOCV R^2	MAE	Pearson r
selected span	27	0.06	0.084	0.30
top-K geometry	27	0.77	0.045	0.88
mean geometry	27	0.75	0.055	0.87
stable dict. rank	27	0.59	0.057	0.77
effective dict. rank	27	0.37	0.065	0.61
compact geometry	27	0.89	0.038	0.94
full geometry	27	-453.25	0.557	0.46
dense ceiling	27	0.80	0.031	0.89
compact geom. + dense	27	0.88	0.031	0.94
full geom. + dense	27	-253.84	0.417	0.48
model size	27	-2.00	0.129	-0.42

out-of-decoder-span teachers before the method is applied to model activations. This is the setting where support F1 is directly meaningful, so it anchors the interpretation of the noisier public-harness and cross-family results.

6. Official MIB Public-Harness Evaluation

We executed the official MIB causal-variable-track public-test run on Gemma-2-2B, evaluating the packaged CSD artifacts with the benchmark evaluator and aggregator. We report uncapped public-split aggregates, with capped split-level diagnostics retained only for per-position analysis.

Table 23 reports the original capped per-split, per-token-position diagnostic scores. The uncapped public-harness aggregate gives CSD scores of 0.487 for MCQA and 0.462 for RAVEL on Gemma-2-2B. These are useful but clearly below dense DAS on RAVEL, where the dense teacher reaches 0.733.

As a diagnostic, we also evaluated the rank-32 dense DAS teachers in the same subspace-artifact format with the same public-test harness. This answers whether the stress-split drop is inherent to dense DAS or introduced by sparse SAE distillation. The uncapped dense-DAS aggregate scores are 0.570 for MCQA and 0.733 for RAVEL. Table 24 compares target positions directly for the earlier capped diagnostic. MCQA remains stress-fragile even for dense DAS, although dense DAS improves the random-letter splits. RAVEL is different: at 2B, CSD reaches only 63% of the dense-DAS aggregate. Thus the 2B RAVEL gap is a sparse-decomposability/generalization gap, not merely a broken stress split.

We then run the smallest cross-scale check on Gemma-2-9B using the same sites and comparable 9B Gemma Scope residual SAEs at layers 10 and 14. A feasibility check confirms that Gemma-2-9B, both 16k residual SAEs, and a forward pass fit on a single high-memory GPU. We train the same rank-32 dense DAS teacher and sparse CSD student, then run the same MIB public-test harness. Table 4 summarizes the uncapped aggregate result. Gemma-2-9B closes the RAVEL SAE-coordinate recovery gap across five seeds: mean CSD is 0.861 versus 0.861 for dense DAS, giving a mean CSD/dense ratio of 1.00. MCQA remains mixed: CSD falls from 0.487 at 2B to 0.422 at 9B, while dense DAS is roughly stable (0.570 at 2B versus a five-seed 9B mean of 0.552). Thus the positive result is specific to RAVEL SAE-coordinate behavioral recovery, not a blanket improvement across causal variables. A targeted MCQA stress diagnostic clarifies the failure mode: when the 9B teacher is trained directly on random-letter counterfactuals, CSD recovers 0.88 against dense DAS 1.00 on the matching random-letter split, but dense DAS itself collapses to 0.00 on the mismatched answer-position and mixed stress splits. This supports reading MCQA as an objective/site/task brittleness case rather than a simple CSD-only failure. Layer-sweep diagnostics give the same qualitative RAVEL conclusion, so the cross-scale finding is not isolated to a single training run. The same stress diagnosis strengthens in the cross-family Llama follow-up: training dense DAS and CSD directly on random-letter counterfactuals gives mean target CSD 0.904

versus dense DAS 0.93 across five seeds and two layers (Table 7), while the mixed answer-position+random-letter objective reaches only mean CSD 0.450 versus dense DAS 0.576. This supports a diagnostic reading: factor-matched format stress can be sparse-recovered, but compositional stress remains harder. Table 8 is the compact scale result: the clean scale claim is 2B→9B RAVEL, while the 27B point is explicitly qualified as a feasibility check with a weak dense-DAS ceiling. Additional layer, width, and stress follow-ups are in Section B. They support the same diagnostic reading rather than a monotonic scaling law. The broader 9B RAVEL picture is layer-sensitive: across off-canonical layers 6, 8, 12, 14, 16, 18, and 21, CSD/dense-DAS ratios range from 0.66 to 0.92, with a seed-2 repeat at layer 10 giving 0.70 and a same-site 131k-width ablation giving 0.72. Thus the 1.00 uncapped five-seed headline should be read as the canonical-layer official-harness result, not a uniform property of every 9B layer. A new 2B seed-4 official rerun reaches ratio 1.002, so this is evidence for a stable positive tuple, not a claim that every 9B layer dominates every 2B run.

This RAVEL recovery is not selector-specific SAE feature discovery. Same-site random-K controls score 0.862 on the 2B seed-4 rerun and 0.865 on the 9B seed-0 row, slightly above the corresponding CSD scores. We therefore use RAVEL as evidence that dense causal behavior can be reproduced in SAE coordinates, but not as evidence that CSD uniquely identifies human-meaningful features at that redundant site. The selector-specific positives are the MCQA rows where matched random-K controls remain small.

Finally, Section 10 gives a preliminary cross-family check. A matched Qwen-Scope probe on Qwen3-1.7B and Qwen3-8B gives the cleanest second-family scale evidence available to us. Across three-seed residual layer grids, Qwen3-1.7B reaches CSD/dense recovery ratios 0.61–0.73, while Qwen3-8B reaches 0.78–0.93. The cleanest 8B positive layer is layer 8 (ratio 0.928 ± 0.017), because layer 24 has a high sparse-recovery ratio but a weaker full-vector ceiling. We therefore treat Qwen as qualified support for the scale-dependence story, not as a second full benchmark. The Llama-3.1-8B check uses EleutherAI residual SAEs. We do not use it as a headline result because the available public SAEs only cover two late residual sites. To avoid treating the Llama result as a post-hoc selector-specific search, we use a fixed gate: first measure full-vector IIA for each candidate layer/position, and only run dense DAS and CSD when the canonical full-vector IIA exceeds 0.70. Table 3 summarizes the outcome. MCQA becomes sparse-recoverable only after the intervention site is moved from the answer symbol to the last prompt token. RAVEL and an additional MLP-site check both show high sparse recovery only relative to a weak dense-DAS ceiling. Repeating the positive last-token MCQA run with seed 1 exactly preserves the canonical dense/CSD scores at both layers.

Table 3. Site validity before sparse recovery on Llama-3.1-8B MCQA. We first test canonical full-vector IIA, then run dense DAS/CSD only where the site is valid. Values show layers 23/29 for seed 0. Seed 1 reproduces the last-token canonical dense/CSD scores.

Position	Full-vector IIA	Decision	CSD after gate
answer symbol	0.04 / 0.04	skip	–
last prompt token	0.92 / 0.92	run	0.94 / 0.92
answer-line token	0.04 / 0.04	skip	–

Table 4. Uncapped scores from our CSD artifacts evaluated with the official MIB causal-variable-track public-harness scripts. MIB is cited for the benchmark, task definitions, evaluator, and public-harness protocol, not as the source of these numeric scores. Ratios are CSD aggregate divided by dense-DAS aggregate.

Model	Task	CSD	Dense DAS	CSD/Dense
Gemma-2-2B	MCQA	0.487	0.570	0.854
Gemma-2-2B	RAVEL	0.462	0.733	0.630
Gemma-2-2B seed 4	RAVEL	0.841	0.839	1.002
Gemma-2-9B seed 0	MCQA	0.425	0.560	0.760
Gemma-2-9B seed 0	RAVEL	0.861	0.860	1.002
Gemma-2-9B seed 1	MCQA	0.425	0.575	0.740
Gemma-2-9B seed 1	RAVEL	0.862	0.863	0.999
Gemma-2-9B seed 2	MCQA	0.425	0.578	0.736
Gemma-2-9B seed 2	RAVEL	0.860	0.859	1.002
Gemma-2-9B seed 7	MCQA	0.417	0.481	0.867
Gemma-2-9B seed 7	RAVEL	0.859	0.864	0.993
Gemma-2-9B seed 8	MCQA	0.417	0.564	0.739
Gemma-2-9B seed 8	RAVEL	0.861	0.860	1.002
Gemma-2-9B mean	MCQA	0.422	0.552	0.768
Gemma-2-9B mean	RAVEL	0.861	0.861	1.000

The layer-sensitive 9B follow-ups reconcile Table 8 with Table 25. The uncapped official-harness 9B five-seed mean ratio is 1.00 at the canonical task layer, while the broader single-seed layer sweep gives ratios 0.66–0.92. We therefore claim that sparse decomposability can be high at valid task sites, not that all larger-model layers are equally sparse-decomposable.

7. Feature Audit

The qualitative audit supports the frontier framing but does not justify a blanket interpretability claim. The important distinction is between causal usefulness and semantic cleanliness. A feature can help reproduce the teacher intervention while still being labeled as a dataset artifact or background correlate. Conversely, a semantically neat feature is not guaranteed to be causally sufficient. Several MCQA features are answer-letter or answer-format related, and several RAVEL features are geography related, but many selected latents appear to be dataset-background artifacts. Tables 30–32 give the audited feature lists, top-activation contexts, and rater breakdown.

A concrete example illustrates the distinction. MCQA feature 47 is labeled as references to the letter “B” and fires

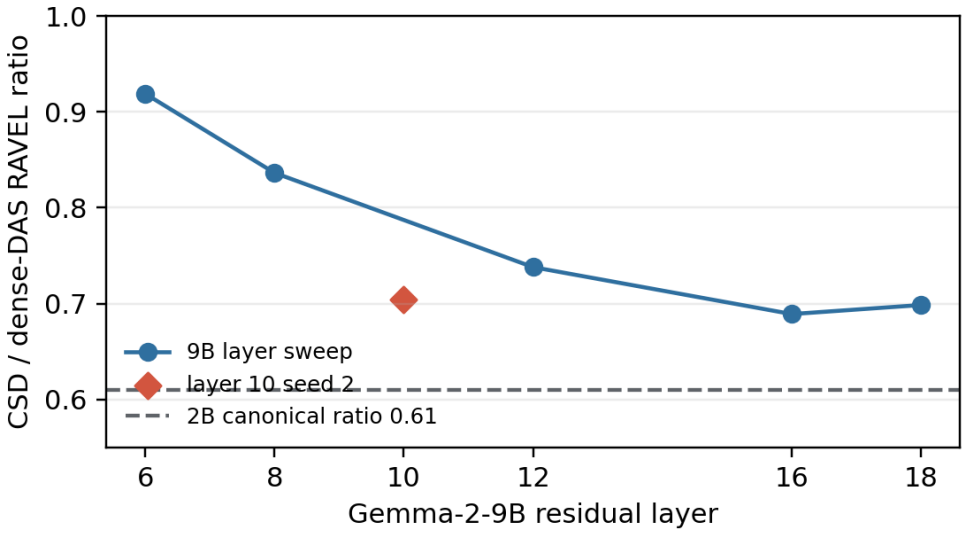

Figure 2. 9B RAVEL layer sweep follow-up. The dashed line is the 2B canonical-layer CSD/dense ratio. The 9B layer sweep is a single-seed follow-up over off-canonical layers, not the official-harness five-seed canonical-layer result. It shows layer sensitivity rather than a uniform 9B scaling law.

Table 5. Earlier Gemma-2-9B capped split-level diagnostic (dataset cap 100). Scores are at the target position used for each task. The uncapped aggregate result used for the headline is Table 4.

Task	Split	CSD	Dense DAS
MCQA	answerPosition_test	0.87	0.96
MCQA	randomLetter_test	0.23	0.38
MCQA	answerPosition_randomLetter_test	0.16	0.34
RAVEL	prompt_template_test	0.75	0.80
RAVEL	attribute_test	0.88	0.91
RAVEL	wikipedia_test	0.79	0.81

Aggregator headline IIA: CSD $(0.42, 0.81)$ vs. dense DAS $(0.56, 0.84)$ for (MCQA,RAVEL).

Table 6. Gemma-2-9B MCQA random-letter-trained stress diagnosis. The matching random-letter split is sparse-recoverable, while mismatched objectives fail at the dense-DAS stage. This table is a diagnostic, not a new MCQA headline.

Eval split	Full-vector	Dense DAS	CSD	Diagnosis
answerPosition	0.82	0.00	0.58	dense objective mismatch
randomLetter	0.48	1.00	0.88	CSD recovers most of dense objective
answerPosition+randomLetter	0.38	0.00	0.16	invalid/weak stress site

on contexts such as “Physical Review B.” This is causally useful for an answer-letter task but not a satisfying semantic explanation of answer choice. In contrast, RAVEL feature 11051 is labeled as names and geographical locations and fires on contexts such as “Oise, France,” which is closer to the continent variable. The audit therefore does not ask whether selected features are useful. The interventions already test that. It asks whether the useful features are semantically clean.

Table 7. Stress-matched Llama-3.1-8B MCQA runs. Dense DAS and CSD are trained on the same counterfactual family used for the target split, using seeds 0–4, layers 23/29, and the last-prompt-token site. Random-letter stress remains sparse-recoverable over ten target rows, while the mixed objective remains only partially recoverable.

Training / target split	Dense DAS	CSD	CSD/Dense
randomLetter	0.930	0.904	0.972
answerPosition+randomLetter	0.576	0.450	0.787

Table 8. RAVEL scale hook. The headline is the official-harness 2B→9B jump in SAE-coordinate behavioral recovery. Matched random-K controls determine whether this is selector-specific. The 27B dense-rescue row is included as qualified site-validity and sparse-limitation evidence rather than as a positive monotonic scale point.

Model	Dense DAS	CSD	CSD/Dense
Gemma-2-2B uncapped official harness	0.733	0.462	0.630
Gemma-2-2B seed-4 official rerun	0.839	0.841	1.002
Gemma-2-9B uncapped official harness, five-seed mean	0.861	0.861	1.000
Gemma-2-27B feasibility, weak dense ceiling	0.29	0.28	0.98
Gemma-2-27B dense-rescued site, sparse-limited	0.667	0.280	0.420

8. Limitations and Next Steps

The current results have six main limitations. First, we report public-harness artifact evaluations rather than private-leaderboard scores (Section 6). Second, cross-family evidence is still feasibility-scale: the Qwen-Scope probe is a three-seed layer grid but not an official MIB-harness run, and the Llama-3.1-8B check in Section 10 uses public EleutherAI SAEs at only the available sites rather than a full matched SAE sweep. Third, MCQA stress robustness is

Table 9. Audit takeaway: causal recovery and semantic plausibility are separate axes. Majority-plausible counts are from the conservative three-rater audit over the five displayed top features per task.

Task	Causal evidence	Semantic audit	Takeaway
MCQA	canonical IIA high (2B 0.96, 9B 0.87)	2/5 displayed rows majority-plausible	causal but artifact-heavy
RAVEL	9B target splits 0.75/0.88/0.79	2/5 displayed rows majority-plausible	more semantic signal
Overall	sparse recovery varies by task/scale	4/10 displayed rows majority-plausible	no blanket naming claim

not solved uniformly: factor-matched random-letter stress is sparse-recoverable in Llama, but the compositional answer-position+random-letter objective remains only partially recoverable. Fourth, the 27B and 131k-width follow-ups are deliberately small: 27B establishes workflow feasibility and yields a dense-valid layer-22/entity-token site, but the bounded CSD-budget sweep stays at DBM-SAE-level recovery. The 9B 131k-width ablation likewise shows that a wider dictionary alone does not guarantee better sparse recovery. Fifth, the semantic audit is mixed, so selected latents should not be read as uniformly human-readable explanations. Sixth, the compact decoder-geometry predictor is still small-N and unbalanced across model families. We therefore treat the leave-one-out result as a useful forecast for these Gemma/Qwen tuples, not as a universal scaling law.

The interpretability audit is also deliberately conservative. CSD can find causally useful latents whose Neuronpedia labels are only weakly related to the target variable. This weakens any simple claim that the selected set is fully human-readable, but strengthens the measurement claim: sparse causal recovery and semantic cleanliness are separate axes.

Taken together, these limitations are part of the diagnostic result: CSD measures how much of a DAS-localized causal variable is recoverable in a pretrained SAE basis, and failures mark where that basis is insufficient.

9. Cross-Family Qwen-Scope Probe

As a matched second-family scale probe, we use Qwen-Scope residual TopK SAEs for Qwen3-1.7B-Base and Qwen3-8B-Base. The available setup is not run through the official MIB harness, but it uses the same internal RAVEL counterfactual splits, rank-32 dense teacher, CSD student, train cap 48, and eval cap 100 as the reviewer follow-ups. We run three seeds over four residual layers per scale: layers 6/10/14/18 for 1.7B and layers 8/12/18/24 for 8B. This is a stronger probe than the earlier one-layer pair, but still not a full cross-family benchmark because the Qwen runs use internal RAVEL splits rather than the public MIB evaluator.

The Qwen result is directionally aligned with the Gemma scale result: sparse recovery improves at the larger checkpoint. It is also more conservative than the Gemma headline: the 8B dense teacher is only moderate, and the 8B layer with the highest ratio is not the layer with the best full-vector ceiling. We therefore use Qwen as evidence that the phenomenon is not obviously Gemma-only, while leaving a full matched cross-family MIB-harness sweep to future work.

Table 10. Matched Qwen-Scope RAVEL layer-grid probe. Values are means over three seeds and RAVEL evaluation splits. The CSD/dense coordinate-recovery ratio improves from Qwen3-1.7B to Qwen3-8B at the best valid sites, but dense-DAS ceilings are moderate, so we report this as qualified cross-family scale evidence.

Model / layer	Full-vector	Dense DAS	CSD	CSD/Dense
Qwen3-1.7B L6	0.773 ± 0.000	0.529 ± 0.007	0.383 ± 0.000	0.725 ± 0.009
Qwen3-1.7B L10	0.748 ± 0.004	0.626 ± 0.004	0.382 ± 0.002	0.611 ± 0.007
Qwen3-1.7B L14	0.707 ± 0.000	0.606 ± 0.005	0.382 ± 0.002	0.631 ± 0.008
Qwen3-1.7B L18	0.637 ± 0.006	0.600 ± 0.020	0.382 ± 0.002	0.638 ± 0.024
Qwen3-8B L8	0.687 ± 0.000	0.449 ± 0.008	0.417 ± 0.000	0.928 ± 0.017
Qwen3-8B L12	0.700 ± 0.000	0.529 ± 0.010	0.417 ± 0.000	0.788 ± 0.015
Qwen3-8B L18	0.693 ± 0.000	0.534 ± 0.015	0.417 ± 0.000	0.780 ± 0.022
Qwen3-8B L24	0.460 ± 0.000	0.452 ± 0.004	0.420 ± 0.000	0.929 ± 0.008

Takeaway: layer 8 is the cleanest 8B positive result, while layer 24 has high ratio but weak full-vector IIA.

10. Cross-Family Llama Feasibility

We ran a small Llama-3.1-8B feasibility check using the public EleutherAI 64x residual SAEs for Llama-3.1-8B at the two available residual hookpoints, `layers.23` and `layers.29` ($d_{\text{SAE}} = 262{,}144$). This is not a full cross-family benchmark: it is a diagnostic check of whether the Gemma conclusions are plausibly specific to Gemma Scope or whether the same site-validity and sparse-recovery phenomena appear with another model/SAE family.

Table 11 shows the MCQA site sweep. The earlier answer-symbol site is invalid on the canonical split: full-vector IIA is only 0.04 at both layers. Moving the intervention to the last prompt token changes the result. At that site, full-vector patching reaches 0.92 on the canonical split, dense DAS reaches 0.92–0.94, and CSD reaches 0.92–0.94. Thus the earlier Llama MCQA failure was primarily a site choice failure, not evidence that Llama MCQA is non-decomposable. A seed-1 repeat of the two positive last-token sites exactly reproduces the canonical dense/CSD scores, while the stress splits remain weaker, consistent with the Gemma story.

Table 11. Llama-3.1-8B MCQA position sweep with EleutherAI 64x residual SAEs (train cap 48, eval cap 50). Dense DAS and CSD are only run when the canonical full-vector IIA exceeds 0.70. Values are seed 0. A seed-1 repeat exactly reproduces the two last-token canonical dense/CSD scores.

Layer	Position	Canonical full vector	Decision	Canonical dense DAS	Canonical CSD
23	answer symbol	0.04	skip	–	–
23	last prompt token	0.92	run	0.92	0.94
23	answer-line token	0.04	skip	–	–
29	answer symbol	0.04	skip	–	–
29	last prompt token	0.92	run	0.94	0.92
29	answer-line token	0.04	skip	–	–

Table 12 reports the RAVEL cross-family feasibility result. The dense-DAS/full-vector ceiling is weak at these Llama SAE sites, but CSD recovers most of the dense signal. Layer

23 has mean CSD/dense recovery 0.86, and layer 29 has mean recovery 0.95. We therefore treat this as support for the diagnostic framing, not as a leaderboard-style improvement: the sparse SAE patch can preserve most of the available dense signal, but the available Llama site does not create a strong dense causal intervention in the first place.

Table 12. Llama-3.1-8B RAVEL feasibility check with EleutherAI 64x residual SAEs (train cap 48, eval cap 100).

Layer	Split	Dense DAS	CSD	CSD/Dense
23	prompt_template_test	0.38	0.33	0.87
23	attribute_test	0.37	0.31	0.84
23	wikipedia_test	0.37	0.32	0.86
29	prompt_template_test	0.35	0.33	0.94
29	attribute_test	0.33	0.31	0.94
29	wikipedia_test	0.33	0.32	0.97

We also tried the available EleutherAI MLP SAEs at layers 23 and 29 as a stronger-site search. This did not improve the dense ceiling: full-vector and dense-DAS RAVEL means stayed near 0.32 at both layers, with CSD matching that weak signal. The result is useful as a negative site search, not as stronger cross-family evidence.

Synthesis. Three points summarize the diagnostic interpretation. First, CSD should be judged against the dense causal teacher, matched random-K controls, and site-validity gates at the same layer and token position. A high CSD score is not meaningful if random-K reaches the same value. In those cases the result diagnoses a redundant or weak benchmark site rather than selector-specific sparse recovery. Conversely, the MCQA positives on Gemma, Llama, and Qwen are useful because the random-K null is small at the same validated site, so the sparse recovery is attributable to the selected SAE subset rather than to arbitrary latent movement.

Second, the negative cases are evidence rather than cleanup material. IOI shows that moving SAE reconstruction error can recover behavior even when named latents do not, arithmetic shows that the current first-answer-token proxy is too weak for a clean sparse-decomposability claim, and 27B RAVEL separates dense-site validity from sparse recovery. These cases prevent the paper from claiming that SAE features are automatically causal explanations. They support the narrower and stronger claim that sparse decomposability is a measured property of a model–site–dictionary tuple.

Third, the feature audit explains how to read the selected latents. The intervention results establish causal usefulness. The audit asks a different question, namely whether useful latents are semantically clean enough for a human-readable circuit story. The answer is mixed, which is why the paper recommends reporting sparse decomposability before naming SAE latents as causal explanations.

11. Conclusion

CSD is best understood as a diagnostic for sparse decomposability. It shows that some dense DAS subspaces are recoverable by a small set of pretrained SAE latents, that scale can change this recoverability, and that causal usefulness is not the same as semantic cleanliness. The failures are therefore not incidental: they identify where the current SAE basis does not provide a faithful sparse description of the causal variable.

References

Arad, D., Mueller, A., and Belinkov, Y. Saes are good for steering – if you select the right features. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10241–10259. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.519.

Bills, S. et al. Language models can explain neurons in language models. OpenAI, project page, 2023. Accessed 2026-05-08.

Bricken, T. et al. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023.

Chan, L., Garriga-Alonso, A., Goldowsky-Dill, N., Greenblatt, R., Nitishinskaya, E., Radhakrishnan, A., and Shlegeris, B. Causal scrubbing: A method for rigorously testing interpretability hypotheses. In *NeurIPS ML Safety Workshop*, 2022.

Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, 2023.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024.

EleutherAI. sae-llama-3.1-8b-64x. Hugging Face, model card, 2024. Accessed 2026-05-08.

Elhage, N. et al. Toy models of superposition. Transformer Circuits Thread, 2022.

Geiger, A., Wu, Z., Potts, C., Icard, T., and Goodman, N. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187, 2024.

Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramar, J., Dragan, A., Shah, R., and Nanda, N. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop*, pp. 278–300. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.19.

Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., and Mueller, A. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations*, 2025.

Mueller, A. et al. Mib: A mechanistic interpretability benchmark. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45069–45108, 2025.

Neuronpedia. Neuronpedia: Interactive reference for sparse autoencoder features. Neuronpedia, 2023. Accessed 2026-05-08.

Qwen Team. Qwen-scope: Turning sparse features into development tools for large language models. Technical report / project release, project page, 2026. Accessed 2026-05-08.

Templeton, A. et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Anthropic, 2024.

Wang, K. et al. Interpretability in the wild: A circuit for indirect object identification in gpt-2 small. In *International Conference on Learning Representations*, 2023.

Wu, Z., Arora, A., Geiger, A., Wang, Z., Huang, J., Jurafsky, D., Manning, C. D., and Potts, C. Axbench: Steering llms? even simple baselines outperform sparse autoencoders. In *International Conference on Machine Learning*, 2025.

A. Validation Controls and Harness Diagnostics

Table 13. Current task-level validation summary. IOI's best score requires the explicit SAE reconstruction-error delta and is therefore diagnostic rather than a named-SAE-only result.

Task	Full	DBM@64	CSD+w@64	Best CSD	Interpretation
MCQA	0.92	0.88	1.00	1.00	Selector-specific sparse recovery
IOI	0.99	0.00	0.00	1.00	Residual/error dominated
RAVEL	0.82	0.16	0.70	0.70	SAE-coordinate recovery; selector-specificity checked separately

Table 14. Full Pareto table for the current validation runs.

Task	K	Random	DBM-SAE	CSD-L1	CSD+w	CSD+w+err	Full
MCQA	32	0.00	0.88	0.88	1.00	–	0.92
MCQA	64	0.00	0.88	0.88	1.00	–	0.92
IOI	8	0.00	0.00	0.00	0.00	1.00	0.99
IOI	32	0.00	0.00	0.00	0.00	1.00	0.99
IOI	64	0.00	0.00	0.00	0.00	1.00	0.99
IOI	128	0.00	0.00	0.00	0.07	1.00	0.99
RAVEL	64	0.16	0.16	0.42	0.70	–	0.82

Table 15. Three-seed validation summary for the headline sparse configurations.

Task	Runs	CSD mean	DBM-SAE mean	Random mean
MCQA	3	1.000	0.880	0.000
RAVEL	3	0.713	0.160	0.160

Table 16. Additional width and teacher-rank ablations. The 65k SAE does not automatically dominate 16k. RAVEL benefits more from a stronger rank-32 teacher than from wider SAE width in this run.

Task	Ablation	Rank	Width	K	CSD	DBM
MCQA	A3_width_65k	8	65k	32	1.000	0.960
RAVEL	A3_width_65k	8	65k	64	0.610	0.170
MCQA	A5_rank_1	1	16k	32	0.980	0.880
MCQA	A5_rank_32	32	16k	32	1.000	0.880
RAVEL	A5_rank_1	1	16k	64	0.680	0.160
RAVEL	A5_rank_32	32	16k	64	0.720	0.160

B. Follow-Up Diagnostics

The follow-ups refine the headline rather than replacing it. Off-headline 9B RAVEL layers have adequate full-vector IIA but weaker dense-DAS ceilings, and CSD tracks that weak dense signal. The same-site 131k-width SAE also fails to improve recovery, so dictionary width alone is not sufficient. The 27B runs show that larger-scale Gemma Scope SAEs can be used in this workflow and that the valid dense site can depend on teacher rank, training budget, and token site: layer 22 at the entity-token site clears the dense gate under the rescue setting, while final-token probes are invalid. The CSD follow-up recovers only 0.28 IIA, matching DBM-SAE and yielding 0.42 of the dense-DAS score, so we treat 27B as dense-valid but sparse-limited rather than as a third monotonic sparse-recovery point. We omit the

Table 17. A4 site ablation comparing residual-stream, MLP-out, and attention Gemma Scope SAEs at the same task layer and headline K.

Task	Site	K	CSD	DBM-SAE	Teacher
MCQA	resid	32	1.000	0.880	1.000
MCQA	mlp	32	0.640	0.020	1.000
MCQA	attn	32	0.000	0.000	0.120
RAVEL	resid	64	0.740	0.160	0.680
RAVEL	mlp	64	0.170	0.160	0.200
RAVEL	attn	64	0.170	0.160	0.170

Table 18. Qualitative sparse-intervention examples. Base is the unpatched base prediction, Source is the counterfactual target answer, and Patched is the model prediction after applying the weighted SAE feature patch.

Task	Base	Source	Patched	Item
MCQA	D	C	C	110
MCQA	C	B	B	111
MCQA	C	B	B	112
RAVEL	Africa	Europe	Europe	Gaya
RAVEL	North America	Europe	Europe	Verona
RAVEL	Africa	Asia	Asia	Malindi

Table 19. Held-out split evaluation using the already-selected headline sparse patches.

Task	Split	N	K	CSD	DBM
MCQA	test	50	32	0.960	0.880
RAVEL	test	100	64	0.780	0.280

Table 20. Three-seed held-out Pareto evaluation for the headline sparse configurations.

Task	Split	K	Runs	CSD mean	CSD std	DBM mean
MCQA	test	8	3	0.853	0.012	0.760
MCQA	test	32	3	0.987	0.023	0.880
MCQA	test	64	3	0.993	0.012	0.880
RAVEL	test	8	3	0.375	0.014	0.248
RAVEL	test	32	3	0.580	0.026	0.242
RAVEL	test	64	3	0.739	0.022	0.242

Table 21. IOI SAE reconstruction-error diagnostic at layer 24, final token.

Patch variant	IIA
Full vector	0.99
All SAE latents + base error	0.01
All SAE latents + source error	0.99
Base latents + source error	0.95
Error delta only	0.95

Table 22. Ablations and controls. Arithmetic is intentionally reported as a negative-control result: under the current binary first-answer-token proxy, even full-vector and teacher interventions are weak.

Ablation	Task	Setting	Source	IIA	Δ
A1_error_term	MCQA	base SAE residual carryover	mcqa_day2_pareto.json	1.00	0.08
A1_error_term	IOI	named SAE features only	ioi_day3_pareto.json	0.00	-0.99
A1_error_term	IOI	source SAE reconstruction error delta	ioi_day3_pareto.json	1.00	0.01
A1_error_term	RAVEL	base SAE residual carryover	ravel_day3_pareto.json	0.70	-0.12
A2_lambda_emb	synthetic	0.0	synthetic_ablations.json	1.00	0.00
A2_lambda_emb	synthetic	0.25	synthetic_ablations.json	1.00	0.00
A2_lambda_emb	synthetic	1.0	synthetic_ablations.json	1.00	0.00
A4_site	IOI	best layer 24 position final_token	ioi_day3_layer_sweep.json	0.99	0.62
A4_site	RAVEL	best layer 22 position final_token	ravel_day3_layer_sweep.json	0.84	0.35
A7_random	MCQA	K=32	mcqa_day2_pareto.json	0.00	-1.00
A8_activation_dbm	MCQA	K=32	mcqa_day2_pareto.json	0.88	-0.12
A7_random	MCQA	K=64	mcqa_day2_pareto.json	0.00	-1.00
A8_activation_dbm	MCQA	K=64	mcqa_day2_pareto.json	0.88	-0.12
A7_random	IOI	K=8	ioi_day3_pareto.json	0.00	0.00
A8_activation_dbm	IOI	K=8	ioi_day3_pareto.json	0.00	0.00
A7_random	IOI	K=32	ioi_day3_pareto.json	0.00	0.00
A8_activation_dbm	IOI	K=32	ioi_day3_pareto.json	0.00	0.00
A7_random	IOI	K=64	ioi_day3_pareto.json	0.00	0.00
A8_activation_dbm	IOI	K=64	ioi_day3_pareto.json	0.00	0.00
A7_random	IOI	K=128	ioi_day3_pareto.json	0.00	-0.07
A8_activation_dbm	IOI	K=128	ioi_day3_pareto.json	0.00	-0.07
A7_random	RAVEL	K=64	ravel_day3_pareto.json	0.16	-0.54
A8_activation_dbm	RAVEL	K=64	ravel_day3_pareto.json	0.16	-0.54
A7_random	ARITHMETIC	K=8	arithmetic_day6_pareto.json	0.00	-0.19
A8_activation_dbm	ARITHMETIC	K=8	arithmetic_day6_pareto.json	0.00	-0.19
A7_random	ARITHMETIC	K=32	arithmetic_day6_pareto.json	0.00	-0.27
A8_activation_dbm	ARITHMETIC	K=32	arithmetic_day6_pareto.json	0.00	-0.27
A7_random	ARITHMETIC	K=64	arithmetic_day6_pareto.json	0.00	-0.27

Table 23. Our CSD artifact evaluated with the official MIB public-test harness on Gemma-2-2B, dataset cap 200. Boldface marks the canonical split for each task (the one the artifact was trained for). Headline is the harness aggregator's per-layer-max over token positions averaged across splits.

Task	Split	correct_symbol	last_token	entity_last_token
MCQA	**answerPosition_test**	**0.96**	0.00	0.00
MCQA	randomLetter_test	0.18	1.00	1.00
MCQA	answerPosition_randomLetter_test	0.35	0.00	0.00
RAVEL	**prompt_template_test**	**0.78**	0.78	0.43
RAVEL	attribute_test	0.52	0.52	0.36
RAVEL	wikipedia_test	0.01	0.01	0.43
Aggregator headline IIA: MCQA $= 0.50$, RAVEL $= 0.44$.				

Table 24. Dense-DAS diagnostic from our artifacts evaluated on the same official MIB public-test harness (dataset cap 200). Scores are at the target position used for each task: `correct_symbol` for MCQA and `entity_last_token` for RAVEL.

Task	Split	CSD	Dense DAS
MCQA	answerPosition_test	0.96	0.96
MCQA	randomLetter_test	0.18	0.26
MCQA	answerPosition_randomLetter_test	0.35	0.51
RAVEL	prompt_template_test	0.43	0.72
RAVEL	attribute_test	0.36	0.70
RAVEL	wikipedia_test	0.43	0.71
Aggregator headline IIA: CSD $(0.50, 0.44)$ vs. dense DAS $(0.58, 0.71)$ for (MCQA,RAVEL).			

Table 25. Follow-up experiments. Means average public-test splits under small dataset caps (50–100 examples per split). The 27B points are feasibility and dense-rescue diagnostics, not official-harness runs.

Check	Setting	Full-vector	Dense DAS	CSD / Ratio
9B RAVEL layer sweep	layers 6/8/12/16/18	0.68–0.79	0.32–0.43	0.29–0.30 / 0.69–0.92
9B RAVEL seed 2	layer 10, 16k	0.78	0.42	0.29 / 0.70
9B RAVEL width	layer 10, 131k	0.78	0.41	0.29 / 0.72
9B MCQA mixed stress	layer 14, mixed train	0.55	0.51	0.55 / –
9B MCQA random-letter train	layer 14, random-letter train	0.48	1.00	0.88 / 0.88
27B RAVEL feasibility	layer 10, 131k	0.80	0.29	0.28 / 0.98
27B RAVEL dense rescue + CSD	layer 22, entity token, rank 128	0.71	0.67	0.28 / 0.42

mixed-MCQA CSD/dense ratio because the mixed training objective depresses the dense-DAS comparison point. The useful observation is qualitative. Training directly on random-letter counterfactuals confirms that the matching random-letter objective is sparse-recoverable (CSD/dense 0.88), while canonical and mixed stress transfer fail at the dense teacher stage. Stress robustness therefore remains a limitation, but the failure is not simply that CSD cannot recover a valid dense MCQA objective.

Benchmark uncertainty and baseline coverage. For interpreting the benchmark claims, the most important appendix material is the audit trail behind each claim. Table 26 reports bootstrap intervals over the finest committed result unit available, and Table 27 records the matched baseline coverage on the positive rows and the Llama stress diagnostic.

Table 26. Uncertainty summary for the main positive and diagnostic claims. Intervals are nonparametric bootstrap intervals over committed result units.

Claim	Metric	n	Mean [95% boot. CI]	Unit
Gemma-2-9B RAVEL official harness	CSD	5	0.861 [0.860, 0.862]	seed
Gemma-2-9B RAVEL official harness	Dense DAS	5	0.861 [0.860, 0.863]	seed
Gemma-2-9B RAVEL official harness	CSD/Dense	5	1.000 [0.996, 1.002]	seed
Llama-3.1-8B MCQA random-letter stress	CSD	10	0.904 [0.892, 0.916]	seed/layer row
Llama-3.1-8B MCQA random-letter stress	Dense DAS	10	0.930 [0.920, 0.940]	seed/layer row
Llama-3.1-8B MCQA random-letter stress	CSD/Dense	10	0.972 [0.958, 0.987]	seed/layer row
Llama-3.1-8B MCQA mixed stress	CSD	10	0.450 [0.440, 0.460]	seed/layer row
Llama-3.1-8B MCQA mixed stress	Dense DAS	10	0.576 [0.544, 0.610]	seed/layer row
Llama-3.1-8B MCQA mixed stress	CSD/Dense	10	0.787 [0.742, 0.831]	seed/layer row
Gemma-2-27B dense-rescued RAVEL site	CSD	3	0.280 [0.220, 0.320]	split
Gemma-2-27B dense-rescued RAVEL site	Dense DAS	3	0.667 [0.520, 0.760]	split
Gemma-2-27B dense-rescued RAVEL site	CSD/Dense	3	0.420 [0.417, 0.423]	split
Synthetic sparse positives	CSD support F1	72	0.958 [0.920, 0.986]	synthetic row
Synthetic dense/out-of-span negatives	CSD high-IIA false positive rate	24	0.000 [0.000, 0.000]	synthetic row

Table 27. Baseline coverage on the main positive configurations and the Llama stress diagnostic.

Setting	Method	n	Mean [95% boot. CI]	Notes
Legacy Gemma-2-2B RAVEL validation, layer 10, K=64	CSD-L1-weighted-edit	3	0.713 [0.700, 0.740]	Same model/site/K validation parity from legacy day9 runs; not Gemma-2-9B and not the official public-harness aggregate.
Legacy Gemma-2-2B RAVEL validation, layer 10, K=64	CSD-L1	3	0.447 [0.380, 0.500]	Same model/site/K validation parity from legacy day9 runs; not Gemma-2-9B and not the official public-harness aggregate.
Legacy Gemma-2-2B RAVEL validation, layer 10, K=64	CSD-OMP	3	0.313 [0.290, 0.340]	Same model/site/K validation parity from legacy day9 runs; not Gemma-2-9B and not the official public-harness aggregate.
Legacy Gemma-2-2B RAVEL validation, layer 10, K=64	DBM-SAE	3	0.160 [0.160, 0.160]	Same model/site/K validation parity from legacy day9 runs; not Gemma-2-9B and not the official public-harness aggregate.
Legacy Gemma-2-2B RAVEL validation, layer 10, K=64	random-K	3	0.160 [0.160, 0.160]	Same model/site/K validation parity from legacy day9 runs; not Gemma-2-9B and not the official public-harness aggregate.
Legacy Gemma-2-2B RAVEL validation, layer 10, K=64	random-K-seeded	3	0.160 [0.160, 0.160]	Same model/site/K validation parity from legacy day9 runs; not Gemma-2-9B and not the official public-harness aggregate.
Llama-3.1-8B MCQA random-letter target rows	Dense DAS	10	0.930 [0.920, 0.940]	Exact same target rows; random-K also succeeds, so this is a site/objective diagnostic.
Llama-3.1-8B MCQA random-letter target rows	CSD	10	0.904 [0.892, 0.916]	Exact same target rows; random-K also succeeds, so this is a site/objective diagnostic.
Llama-3.1-8B MCQA random-letter target rows	DBM-SAE	10	0.560 [0.536, 0.584]	Exact same target rows; random-K also succeeds, so this is a site/objective diagnostic.
Llama-3.1-8B MCQA random-letter target rows	activation-topK	10	0.710 [0.680, 0.740]	Exact same target rows; random-K also succeeds, so this is a site/objective diagnostic.
Llama-3.1-8B MCQA random-letter target rows	DiffMean-to-SAE	10	0.720 [0.660, 0.780]	Exact same target rows; random-K also succeeds, so this is a site/objective diagnostic.
Llama-3.1-8B MCQA random-letter target rows	random-K	10	0.920 [0.920, 0.920]	Exact same target rows; random-K also succeeds, so this is a site/objective diagnostic.

Ground-truth calibration and selector controls. The synthetic battery gives the benchmark-style ground truth that the real tasks cannot provide: planted sparse supports are recovered, while dense or out-of-span controls are rejected. The selector-specificity table records which real-task rows beat matched sparse baselines and which rows should be read as random-K degeneracy.

C. Feature Audit Detail

Table 28. Synthetic sparse-decomposability robustness battery copied from the extended experimental appendix.

Cell	Scenario	Full	CSD IIA	CSD F1	F1 margin
01_aligned_k1	aligned_sparse	0.99	0.99	1.00	0.00
02_aligned_k4	aligned_sparse	1.00	1.00	1.00	0.00
03_aligned_k8	aligned_sparse	1.00	1.00	1.00	0.00
04_aligned_k4_d64	aligned_sparse	1.00	1.00	1.00	0.00
05_aligned_k8_d64	aligned_sparse	1.00	1.00	1.00	0.00
06_correlated_mild	correlated_distractors	1.00	1.00	1.00	0.00
07_correlated_strong	correlated_distractors	1.00	1.00	1.00	0.00
08_noise_low	reconstruction_noise	0.99	0.99	1.00	0.00
09_noise_high	reconstruction_noise	0.97	0.97	1.00	0.00
10_dense_rotated_d32	dense_rotated	1.00	0.01	0.40	0.00
11_dense_rotated_d64	dense_rotated	1.00	0.01	0.22	0.00
12_out_of_span_d32	out_of_span	1.00	0.00	0.00	0.00
13_out_of_span_d64	out_of_span	1.00	0.00	0.00	0.00
14_no_embedding_loss	correlated_distractors	1.00	1.00	1.00	0.00
15_no_error_carryover	reconstruction_noise	0.97	0.97	1.00	0.00
16_low_budget	correlated_distractors	1.00	1.00	1.00	0.00

Table 29. Selector-specificity gap from artifacts evaluated with the official MIB public harness.

Row	Dense	CSD	Random-K	DBM	Max sparse	Gap	Diagnosis
MCQA canonical	0.960	0.960	0.000	0.900	0.900	0.060	selector-specific vs random; modest CSD edge over DBM
MCQA aggregate	0.570	0.487	0.333	0.462	0.462	0.025	structured sparse support; small CSD edge
RAVEL 2B seed4	0.839	0.841	0.862	–	0.862	-0.021	random-K degenerate
RAVEL 9B seed0	0.860	0.861	0.865	–	0.865	-0.004	random-K degenerate

Table 30. Initial feature audit for the sparse SAE success cases. All rows use the Gemma Scope release `gemma-scope-2b-pt-res-canonical`.

Task	Layer	SAE id	K	Rank	Feature	Weight	Label
MCQA	14	layer_14/width_16k/canonical	32	1	13373	0.26	financial metrics and conditions in the context of investment...
MCQA	14	layer_14/width_16k/canonical	32	2	47	1.62	references to the letter 'B' or words starting with 'B'
MCQA	14	layer_14/width_16k/canonical	32	3	12895	1.14	terms related to content networking and computer science conc...
MCQA	14	layer_14/width_16k/canonical	32	4	3615	0.48	questions and queries seeking the closest or smallest values...
MCQA	14	layer_14/width_16k/canonical	32	5	8203	1.37	occurrences of the letter 'A' and variations of it in differe...
MCQA	14	layer_14/width_16k/canonical	32	6	13049	1.74	references to scientific papers and their corresponding ident...
MCQA	14	layer_14/width_16k/canonical	32	7	11702	0.17	numerical values and formal references in academic or structu...
MCQA	14	layer_14/width_16k/canonical	32	8	14184	1.83	mathematical notations and symbols
MCQA	14	layer_14/width_16k/canonical	32	9	13558	0.35	elements related to coding structures and variable assignments
MCQA	14	layer_14/width_16k/canonical	32	10	11853	0.17	references to programming elements related to data structures...
RAVEL	10	layer_10/width_16k/canonical	64	1	52	0.68	terms related to technological developments and innovations i...
RAVEL	10	layer_10/width_16k/canonical	64	2	60	0.12	references to statistical analyses and methodologies
RAVEL	10	layer_10/width_16k/canonical	64	3	6	1.26	references to countries and their associated historical or po...
RAVEL	10	layer_10/width_16k/canonical	64	4	3806	1.94	names and terms related to authors and contributors in academ...
RAVEL	10	layer_10/width_16k/canonical	64	5	11051	1.94	names and geographical locations
RAVEL	10	layer_10/width_16k/canonical	64	6	15497	1.87	references to various Asian ethnicities and cultures
RAVEL	10	layer_10/width_16k/canonical	64	7	3669	1.96	terms and concepts related to slavery and African history
RAVEL	10	layer_10/width_16k/canonical	64	8	417	1.71	geographic locations and their associated identifiers
RAVEL	10	layer_10/width_16k/canonical	64	9	4016	1.91	references to European political and cultural terms, particul...
RAVEL	10	layer_10/width_16k/canonical	64	10	13170	1.82	references to specific individuals, particularly in the conte...

Table 31. Qualitative top-activation exemplars for audited features. Contexts are from Neuronpedia activation records. The max-activating token is bracketed.

Task	Rank	Feature	Label	Top activation context
MCQA	1	13373	financial metrics and conditions in the context of...	\n\n c . The constant growth model cannot be used because the growth rate...
MCQA	2	47	references to the letter 'B' or words starting with...	, Lu Z Y and Zhang Q M 2 0 0 9 [* Physical Review [B] *] [** 7 9 **]...
MCQA	3	12895	terms related to content networking and computer sc...	an Internet Protocol network . \n Related Art \n Recent research efforts...
MCQA	4	3615	questions and queries seeking the closest or smalle...	5 * u = - 2 6 . Which is the third smallest value ? ([a]) - 3 (b) 0 (...
RAVEL	1	52	terms related to technological developments and inn...	Nevada \n\n\n\n\n\n\n\n App ellate Case Management System \n\n\n\n\n C -...
RAVEL	2	60	references to statistical analyses and methodologies	before 2 0 0 3 . 1 9 8 7 Truth about AIDS \n\n Life [expectancy] will go...
RAVEL	3	6	references to countries and their associated histor...	spend my free time writing and reading . My favorite authors are Mindy K...
RAVEL	4	3806	names and terms related to authors and contributors...	corpse of a woman who died in a house at the bottom . Her name was Rachel...

Table 32. Three-rater LLM audit of whether top features match the target causal variable. The displayed audit is conservative: 4 of 10 rows receive a majority-plausible rating, so the current interpretability claim is mixed rather than uniform.

Task	Category	Rank	Feature	Mean	Consensus
MCQA	artifact	1	13373	0.67	no
MCQA	answer_letter	2	47	3.00	yes
MCQA	artifact	3	12895	0.67	no
MCQA	unclear	4	3615	2.00	yes
MCQA	artifact	5	8203	0.67	no
RAVEL	artifact	1	52	0.67	no
RAVEL	background	2	60	0.00	no
RAVEL	geography	3	6	2.33	yes
RAVEL	artifact	4	3806	0.00	no
RAVEL	geography	5	11051	2.67	yes