Geometry, Not Scale Alone, Predicts Sparse Recovery of Causal Subspaces

Joy Zheyun Yang* JOY@ROBOTS.OX.AC.UK
University of Oxford, Oxford, United Kingdom

Socrates Osorio* SOCRATESJ.OSORIO@BERKELEY.EDU
Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA, USA

Abstract

Whether high-dimensional causal representations become easier to recover in sparse bases is not determined by model size alone. We study sparse decomposability: the extent to which a dense DAS-localized causal subspace can be expressed by a small set of pretrained SAE latents at a particular model–site–dictionary tuple, measured by causal sparse distillation (CSD), which matches a dense causal teacher with an SAE-constrained student. The central finding is geometric: two compact pre-CSD decoder statistics, the top-K decoder-subspace cosine with the teacher and the decoder stable rank, predict CSD/dense recovery across dense-valid Gemma/Qwen tuples with leave-one-out $R^2 = 0.89$, leave-one-model-out $R^2 = 0.80$, and bootstrap 95% CI $[0.79, 0.95]$, while model size alone gives $R^2 = -2.00$. This explains why some larger-model sites recover cleanly, others are random-K degenerate, and a 27B dense-valid site remains sparse-limited. A ground-truth synthetic battery and cross-family checks on Gemma, Llama, and Qwen-Scope calibrate the measurement, and matched random-K controls show that high CSD/dense recovery is not by itself evidence of meaningful feature selection. The result is a high-dimensional learning story about how dictionary geometry, scale, and causal subspace orientation jointly control sparse recovery.

1. Introduction

Mechanistic interpretability has two useful but mismatched tool families. Supervised causal localizers, especially DAS [9], optimize directly for interchange intervention accuracy (IIA) and can identify faithful subspaces, but their outputs 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 [12]. If sparse decomposability itself changes with scale or intervention site, then neither family's conclusions are 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? We introduce causal sparse distillation (CSD), a teacher–student procedure that selects SAE features by matching the paired counterfactual intervention behavior of a DAS-style teacher, producing a faithfulness–sparsity curve and a set of candidate features for audit. CSD treats a pretrained SAE basis as fixed: a high score means the dense variable is sparse-decomposable in the available dictionary, while a low score is evidence about the mismatch between that dictionary and the causal variable, not merely an optimization failure. We use *tuple* to

. * Equal contribution.

© J.Z. Yang & S. Osorio.

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. *Model size does not by itself determine sparse causal recovery. The controlling object is the geometry of a dense causal subspace relative to a high-dimensional SAE decoder.*

A high-dimensional learning question. The object of study is the geometry of sparse recovery in overcomplete learned neural dictionaries, and how that geometry moves with model scale, placing the paper in the scope of feature learning and representation in scaling regimes: an SAE decoder is a learned, highly overcomplete dictionary in a high-dimensional residual stream, a DAS teacher is a learned low-rank causal subspace in the same space, and sparse recovery is a question about the relative orientation and effective rank of the two. The contribution speaks to this year's scaling-laws theme in the negative and in the positive. Negatively, parameter count is a poor predictor of sparse causal recovery: across the 2B, 9B, and 27B Gemma checkpoints, recovery is non-monotonic, and a model-size regressor has leave-one-out $R^2 = -2.00$. Positively, on 27 dense-valid Gemma/Qwen tuples, two compact pre-CSD decoder statistics, the top-K decoder-subspace alignment with the teacher and the decoder stable rank, predict CSD/dense recovery with leave-one-out $R^2 = 0.89$. The evidence is therefore a geometry-mediated scale effect, not a scaling law: scale changes recovery only through its effect on dictionary–subspace geometry.

The remaining evidence maps the boundary of this diagnostic result across tasks, sites, and model families. Our strongest behavioral-recovery result is on RAVEL continent interventions, where an SAE-coordinate patch matches dense DAS in a five-seed Gemma-2-9B check on the public MIB harness, but matched random-K controls make it a coordinate-recoverability result rather than selector-specific feature discovery. The selector-specific positives are MCQA sites where matched controls stay low. Appendix Table 3 states the claims we intend the paper to make, and just as importantly the claims it does not make. We do not claim monotonic scaling, leaderboard state of the art, uniformly semantic features, or that CSD always beats random-K: 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 [12]. Gemma Scope provides the frozen pretrained SAE dictionary used here [10]. Sparse Feature Circuits rank SAE features by attribution-style influence on model behavior [11]. Our selection signal instead comes from paired interchange interventions generated by a teacher subspace. SAE motivation builds on superposition and dictionary-learning accounts of polysemantic representations [3, 6, 8, 15], and on causal-intervention circuit work [4, 5, 16]. Feature labels and top-activation exemplars are used only as audit evidence, following automatic neuron-description and Neuronpedia-style workflows [2, 13]. Closest in spirit are supervised feature-selection methods that add task information, including output-effect steering selection [1] and benchmark results showing that simple steering baselines can beat raw SAE features [17]. Those methods optimize scalar output effects or attribution-style rankings, whereas CSD optimizes paired counterfactual agreement with a dense causal teacher and reports how much of the teacher survives at each sparsity level, giving an operational test of whether an opaque causal subspace is expressible in a fixed, named SAE basis.

3. Method

Each training item is 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 teacher-subspace matching:

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

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, alongside 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 [14] and EleutherAI Llama-3.1-8B residual SAEs [7] where noted. The main MIB-style tasks are MCQA answer pointer, IOI, and RAVEL continent, with arithmetic carry 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. We evaluate with the public MIB harness and do not claim a private-test leaderboard result. 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, and 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

5.1. Dictionary geometry predicts sparse recovery

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) and evaluate linear predictors by leave-one-out cross-validation. Table 1 shows that compact geometry predicts sparse recovery while model size alone does not.

Table 1: 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

The geometric mechanism explains the scale pattern. Changing model scale changes the orientation and concentration of causal subspaces relative to the SAE decoder, but it does not guarantee a monotonic increase in sparse recovery: the 9B canonical RAVEL site is highly recoverable in SAE coordinates but random-K degenerate, off-canonical 9B layers are only partially recoverable, and the 27B dense-rescued site is valid for dense intervention while remaining sparse-limited. This is exactly the pattern expected if sparse decomposability is controlled by decoder-subspace alignment and effective rank rather than parameter count: the top-K decoder-subspace cosine measures whether the dictionary has directions aligned with the dense teacher, stable rank measures how diffuse the relevant decoder geometry is, and together they summarize whether a small coordinate subset can span the causal displacement. CSD is the empirical recovery test. The decoder-geometry features are the pre-experiment forecast of when that test should succeed.

5.2. Task-level boundary of sparse decomposability

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 is recoverable only when the SAE reconstruction error moves, not from named SAE latents alone. Arithmetic remains a negative control under the binary first-answer-token proxy. Appendix B gives the detailed controls, seed stability, site ablations, and qualitative examples.

5.3. Official MIB harness self-host and cross-scale checks

We replicate the leaderboard's public-test evaluation pipeline locally using the official MIB causal-variable-track evaluator and aggregator on Gemma-2-2B against the packaged submission artifacts. The uncapped aggregate gives CSD 0.487 for MCQA and 0.462 for RAVEL. Packaging the rank-32 dense DAS teachers themselves as MIB subspace submissions gives dense-DAS aggregates of 0.570 and 0.733. MCQA remains stress-fragile even for dense DAS, while at 2B RAVEL, CSD reaches only 63% of the dense-DAS aggregate, so the 2B RAVEL gap is a sparse-decomposability/generalization gap, not merely a broken stress split (Tables 14 and 15).

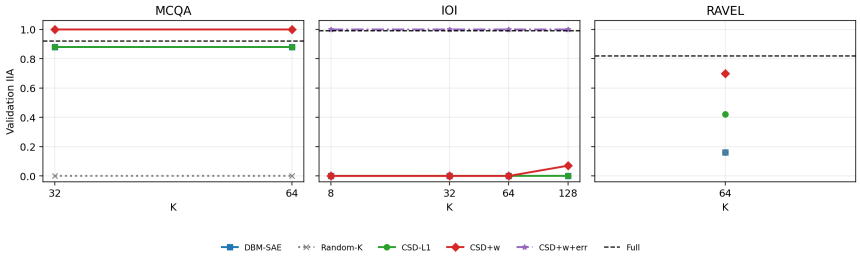

Figure 1: Validation IIA versus selected SAE feature count. MCQA shows selector-specific sparse recovery. RAVEL requires random-K controls to distinguish true feature selection from redundant-site degeneracy. IOI only recovers when the source-error delta is explicitly allowed.

Table 2: RAVEL scale hook. The headline is the official-harness 2B→9B jump in SAE-coordinate recovery. The 27B rows are qualified feasibility and sparse-limitation evidence, not monotonic scale points.

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

We then run the smallest cross-scale check on Gemma-2-9B with matched 9B Gemma Scope residual SAEs at the same task sites and the same teacher/student setup (per-seed results in Table 16). Gemma-2-9B closes the RAVEL SAE-coordinate recovery gap across five seeds: mean CSD is 0.861 versus 0.861 for dense DAS, 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, so the positive result is specific to RAVEL SAE-coordinate behavioral recovery, not a blanket improvement. A targeted MCQA stress diagnostic clarifies the failure mode (Table 18): a 9B teacher trained directly on random-letter counterfactuals gives CSD 0.88 against dense DAS 1.00 on the matching split, but dense DAS itself collapses to 0.00 on the mismatched answer-position and mixed stress splits, supporting an objective/site/task brittleness reading rather than a CSD-only failure. The same diagnosis strengthens in the cross-family Llama follow-up: stress-matched random-letter training gives mean target CSD 0.904 versus dense DAS 0.93 across five seeds and two layers (Table 19), while the mixed answer-position+random-letter objective reaches only 0.450 versus 0.576. Factor-matched format stress can be sparse-recovered, but compositional stress remains harder.

Table 2 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. The broader 9B RAVEL picture is layer-sensitive: across off-canonical layers, CSD/dense-DAS ratios range from 0.66 to 0.92 (Appendix C and Figure 2), so the 1.00 five-seed headline is the canonical-layer official-harness result, not a uniform property of every 9B layer. A 2B seed-4 official rerun reaches ratio 1.002, evidence for a stable positive tuple rather than 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, so RAVEL shows that dense causal behavior can be reproduced in SAE coordinates, not that CSD uniquely identifies human-meaningful features there. The selector-specific positives are the MCQA rows where matched controls remain small.

Finally, the cross-family appendices give preliminary checks. A matched Qwen-Scope probe 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 (Appendix D) — qualified support for the scale-dependence story, not a second full benchmark. The Llama-3.1-8B check (Appendix E) uses a fixed gate to avoid a post-hoc selector-specific search: dense DAS and CSD run only where canonical full-vector IIA exceeds 0.70. The answer-symbol site is invalid (full-vector IIA 0.04), while the last prompt token is a valid site (full-vector 0.92) where CSD again matches dense DAS (0.92–0.94 at layers 23/29, reproduced at seed 1). RAVEL and an MLP-site check show high sparse recovery only relative to a weak dense-DAS ceiling (Table 23).

5.4. Feature audit

The qualitative audit separates causal usefulness from semantic cleanliness: a feature can help reproduce the teacher intervention while still being labeled as a dataset artifact or background correlate. MCQA feature 47, labeled as references to the letter "B" (firing on "Physical Review B"), is causally useful for an answer-letter task but not a satisfying semantic explanation of answer choice, while RAVEL feature 11051, labeled as names and geographical locations, is closer to the continent variable. Across the conservative three-rater audit, 4/10 displayed top features are majority-plausible (Table 20).

6. Limitations and Next Steps

The current results have six main limitations. First, they are not a state-of-the-art leaderboard claim: we self-host the official MIB public-test harness and do not claim a private-leaderboard score. Second, cross-family evidence is still feasibility-scale (the Qwen-Scope probe is not an official MIB-harness submission, and the Llama-3.1-8B check uses public EleutherAI SAEs at only the available sites). Third, MCQA stress robustness is not solved uniformly: the compositional answer-position+random-letter objective remains only partially recoverable. Fourth, the 27B and 131k-width follow-ups are deliberately small, and both show that neither scale nor dictionary width alone guarantees better sparse recovery. Fifth, the semantic audit is mixed. Sixth, the compact decoder-geometry predictor is still small-N and unbalanced across model families: a useful forecast for these Gemma/Qwen tuples, not a universal scaling law.

7. Conclusion

CSD is a diagnostic for sparse decomposability: it measures how much of a DAS-localized causal variable is recoverable in a pretrained SAE basis, and the failures are part of the result. Some dense DAS subspaces are recoverable by a few pretrained SAE latents, scale changes recoverability only through dictionary–subspace geometry, and causal usefulness is not semantic cleanliness: failures mark where the SAE basis cannot faithfully describe the causal variable.

Acknowledgments

In accordance with the workshop's disclosure policy, we disclose that large language models were used as writing and editing aids for parts of the text. All technical content, experiments, and claims were authored and verified by the authors.

References

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

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

[3] Trenton Bricken et al. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023.

[4] Lawrence Chan, Adria Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Ekaterina Nitishinskaya, Ansh Radhakrishnan, and Buck Shlegeris. Causal scrubbing: A method for rigorously testing interpretability hypotheses. In *NeurIPS ML Safety Workshop*, 2022.

[5] Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adria Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, 2023.

[6] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024.

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

[8] Nelson Elhage et al. Toy models of superposition. Transformer Circuits Thread, 2022.

[9] Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 160–187, 2024.

[10] Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop*, pages 278–300. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.19.

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

[12] Aaron Mueller 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*, pages 45069–45108, 2025.

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

[14] 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.

[15] Adly Templeton et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Anthropic, 2024.

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

[17] Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. Axbench: Steering llms? even simple baselines outperform sparse autoencoders. In *International Conference on Machine Learning*, 2025.

Appendix A. Claims and Evidence

Table 3: 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

Appendix B. Validation Controls and Harness Diagnostics

Table 4: 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 5: 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 6: 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 7: 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

Table 8: 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 9: 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 10: 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 11: 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 12: 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 13: 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 14: 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 submission 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 15: Dense-DAS diagnostic from our artifacts evaluated on the same official MIB public-test harness (dataset cap 200). Scores are at the submitted target position 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 16: Uncapped scores from our submitted 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

Table 17: Earlier Gemma-2-9B capped split-level diagnostic (dataset cap 100). Scores are at the submitted target position for each task. The uncapped aggregate result used for the headline is Table 16.

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 18: 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

Table 19: 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. 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 20: 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

Table 21: 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 submissions.

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

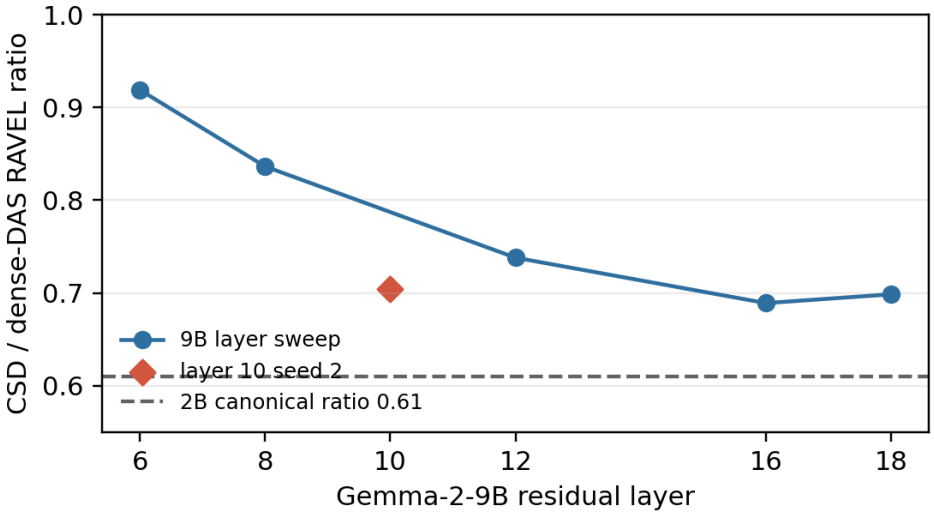

Figure 2: 9B RAVEL layer sweep follow-up. The dashed line is the 2B canonical-layer CSD/dense ratio. The 9B layer sweep is a single-seed follow-up over off-canonical layers, not the official-harness five-seed canonical-layer result. It shows layer sensitivity rather than a uniform 9B scaling law.

Appendix C. 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 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.

Appendix D. Cross-Family Qwen-Scope Probe

We searched for a clean matched second-family scale pair and found Qwen-Scope residual TopK SAEs for Qwen3-1.7B-Base and Qwen3-8B-Base. The available setup is not an official MIB harness submission, 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.

Table 22: 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.

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.

Appendix E. 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 23 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 24 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

Table 23: 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, and 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	–	–

most of the available dense signal, but the available Llama site does not create a strong dense causal intervention in the first place.

Table 24: 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.

Appendix F. Geometry-Predictor and Uncertainty Details

Table 25 reports bootstrap intervals for the main positive and diagnostic rows. The predictor ablation itself is in the main text (Table 1).

Synthetic high-dimensional controls. The synthetic battery separates true sparse decomposition from dense or out-of-span teachers. It is useful for HiLD because it shows that recovery is not a trivial consequence of high-dimensional overcompleteness: when the dense teacher is rotated away from the sparse basis, the high-IIA false-positive rate is zero.

High-dimensional null controls. The same overcomplete dictionary can make random sparse directions look surprisingly strong when the intervention site is redundant. For that reason, we also keep the baseline-parity and selector-gap audits: Table 27 records when CSD is separated from DBM-SAE, activation-topK, DiffMean-to-SAE, and random-K controls, and Table 28 marks rows where high recovery is better read as random-K degeneracy.

Table 25: 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 26: Synthetic sparse-decomposability robustness battery.

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 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-refit	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.
	CSD-L1	3	0.447 [0.380, 0.500]	
	CSD-OMP	3	0.313 [0.290, 0.340]	
	DBM-SAE	3	0.160 [0.160, 0.160]	
	random-K	3	0.160 [0.160, 0.160]	
	random-K-seeded	3	0.160 [0.160, 0.160]	
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.
	CSD	10	0.904 [0.892, 0.916]	
	DBM-SAE	10	0.560 [0.536, 0.584]	
	activation-topK	10	0.710 [0.680, 0.740]	
	DiffMean-to-SAE	10	0.720 [0.660, 0.780]	
	random-K	10	0.920 [0.920, 0.920]	

Table 28: 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