

Dictionaries Without Ontologies: What a Sparse Autoencoder Latent Is Evidence For, Why Decodability, Use and Reproducibility Do Not Add Up to Canonicity, and the Identifiability Premise No Real-Model Result Has Checked

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Sparse autoencoders are now the default instrument for decomposing language model activations, and their latents are routinely reported as the model's features: units the network itself uses, which a wide enough dictionary has pulled out of superposition. This paper asks what a latent is evidence for. It separates four claims that the word "feature" runs together - that a property is linearly readable at an activation site, that the model's downstream computation depends on a direction, that the same unit recurs across training choices, and that the recurring unit is canonical, meaning unique, complete and atomic - and reads the published record against each. The record supports the first broadly, though the natural-language label attached to a latent is a weaker link than the latent itself; it supports the second unevenly, with causal roles that differ between SAE families trained on the same model; it supports the third at the grain of subspaces more readily than at the grain of individual latents; and it supports the fourth on no real language model. The reason is structural. Classical results make sparse dictionaries unique only when the data are generated by a sparse linear code meeting stated conditions; that premise has been tested only piecemeal, concept by concept, and the comparisons with a ground truth for the full dictionary that could reach canonicity have been run only on toy, synthetic, formal-language and board-game models; natural-language models have only task-specific, approximate ground truths. A short rotation argument shows how a multi-dimensional feature would produce the pattern of seed-dependent latents inside reproducible subspaces that has been reported. The paper ends with a use-indexed licensing procedure and five studies that would settle the open part. Throughout, it works from published measurements, each credited to the source that reported it.

Keywords: sparse autoencoders; mechanistic interpretability; dictionary learning; identifiability; superposition; feature universality; causal abstraction; evaluation validity

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

The sparse autoencoder (SAE) programme in mechanistic interpretability rests on a compact argument. Networks appear to represent more properties than they have dimensions by overlapping them in superposition, a phenomenon made precise in toy models by Elhage et al. (2022). If activations are sparse combinations of many directions, a wide dictionary trained to reconstruct them under a sparsity penalty should recover those directions, and each recovered latent should then name something the network represents (Cunningham et al., 2023). The programme has scaled quickly: k-sparse autoencoders with clean scaling laws and a 16-million-latent dictionary trained on GPT-4 activations (Gao et al., 2024), open suites covering every layer of Gemma 2 2B and 9B (Lieberum et al., 2024) and Llama-3.1-8B (He et al., 2024), and a succession of architectures - gated (Rajamanoharan et al., 2024a), JumpReLU (Rajamanoharan et al., 2024b), BatchTopK (Bussmann et al., 2024) and Matryoshka (Bussmann et al., 2025) - each aimed at a better trade-off between reconstruction and sparsity or at better-organised features. Latents are then used as the units of circuit analysis (Marks et al., 2024), the handles for steering and unlearning (Farrell et

al., 2024; Muhamed et al., 2025), and the evidence in alignment audits (Marks et al., 2025).

The operative premise is rarely stated outright, and it is stronger than any of these uses requires. It is that an SAE latent is a discovered fact about the model: that the dictionary recovers the network's own units of representation, as opposed to one convenient basis among many for describing its activations. Paulo and Belrose (2025) state the deflationary alternative plainly: the features an SAE finds should be read as a useful decomposition of activation space rather than an exhaustive and universal list of what the model "truly" uses. Leask et al. (2025) put the realist version in its most testable form - the hope that SAEs return a canonical set of units, unique, complete and atomic - and argue against it. Song et al. (2025) and Nelson et al. (2026) respond from the other direction: instability is real, but it can be reduced by architectural choices, and the sparse codes of dictionaries learned in practice may be close to identifiable.

Both sides cite the same artifacts, publish in the same venues and largely agree on the measurements. The disagreement is about what the measurements license. This paper's contribution is to make that question answerable by splitting it. Section 2 defines four claims,

in ascending order of demand, that the word "feature" is used to make about a latent: that a property is decodable at a site (Tier 1), that the model uses a direction (Tier 2), that the unit is reproducible across training choices (Tier 3), and that it is canonical (Tier 4). It then states what identifiability theory says Tier 4 would require, and gives a short rotation argument, labelled as the author's own, for why multi-dimensional structure in activations would produce seed-dependent latents inside reproducible subspaces. Section 3 asks what the standard proxies - reconstruction, sparsity, automated interpretability - measure, using a control that separates the model from the procedure: an SAE trained on a model whose weights are random. Sections 4 to 7 read the evidence tier by tier. Section 8 gives the strongest case for the programme, which turns out to be a case for uses that do not need Tier 4. Section 9 states a use-indexed licensing procedure (Algorithm 1, Table 2). Section 10 proposes five studies that would settle the open part, and Section 11 states what this paper does not show.

Three things this paper does not argue should be said at the start. It does not argue that SAEs are useless; several of the uses in Section 8 are well-supported. It does not argue that language models have no canonical units; it argues that no published test on a model trained on natural language was built so that it could confirm them if they exist, and that tests which could be specifiable and feasible. And it does not re-argue that SAE latents are unstable across seeds, which Paulo and Belrose (2025), Fel et al. (2025) and Nelson et al. (2026) have established; the instability is an input to the argument, not its conclusion.

The closest prior positions should also be named. Peng et al. (2025) reconcile the positive and negative SAE literatures by separating discovery of unknown concepts from action on known ones; this paper agrees with the separation and adds that discovery about a dataset and discovery about a model make different demands (Section 8). Song et al. (2025) propose feature consistency as the quantity to optimise; this paper treats consistency as Tier 3 evidence and argues that it does not reach Tier 4 without a generative premise (Section 7). Hindupur et al. (2025) show that each SAE architecture projects assumptions about concept geometry onto what it can find; this paper builds on that result directly. The delta here is the tiered decomposition, the observation that the proxies in use sit at identifiable tiers and none above Tier 3, the use-indexing, and the explicit statement of which premise a canonicity claim would have to check.

2. Four claims under one word

2.1 The tiers

Fix a model M , an activation site s (a layer and sub-layer), a dictionary D learned there by some procedure P (architecture, width, sparsity mechanism, data, seed), and a latent d_i in D with an attached natural-language label L_i . The following four claims are different, and each is stipulated here.

Tier 1 - Decodability. The property named by L_i is linearly readable from M 's activations at s : on held-out inputs, the latent's activation predicts the property's presence better than a stated baseline. This is a claim about the information content of the activations. It says nothing about whether M uses that information, and it is satisfied by any direction that a supervised probe could also find.

Tier 2 - Use. M 's downstream computation depends on the direction d_i : intervening on the latent's contribution changes M 's outputs in the way L_i predicts, more than matched interventions on comparison directions do. This is a causal claim about M , indexed to an intervention type (zero-ablation, mean ablation, clamping, additive steering) and an output measure.

Tier 3 - Reproducibility. The unit recurs when P changes in ways the practitioner regards as irrelevant: across seeds, across widths, across sparsity mechanisms, and ideally across related models. Reproducibility is a property of the pair $(M, P\text{-family})$. It can hold at the grain of individual latents or only at the grain of a subspace they jointly span.

Tier 4 - Canonicity. The unit is one of M 's own units of representation: the decomposition is unique up to trivial symmetries (permutation and scaling), complete (nothing the model represents at s is missing), and atomic (no unit is itself a composite of others). This is the claim Leask et al. (2025) formalise and test. It is the claim implied when a paper says that an SAE has found "the features" a model uses, or when the absence of a latent is read as the absence of a representation.

The tiers are ordered by demand but are not a ladder in the logical sense: Tier 2 does not entail Tier 1 (a direction can matter causally without having a clean description of the inputs that activate it, which is the situation Arad et al. (2025) call an output feature), and Tier 3 entails neither of the first two (a procedure can reproducibly find directions that M does not use). What the order captures is the evidence each requires. Tier 1 needs held-out prediction and a baseline. Tier 2 needs interventions and comparison directions. Tier 3 needs replication across P . Tier 4 needs either a ground truth to compare against or a theorem that makes the decomposition unique given a premise about how the activations were generated, plus a check of that premise.

2.2 The label as a separate claim

Most published uses of a latent pass through its label. Labels are now produced at scale by automated pipelines that ask one language model to describe a latent's top-activating contexts and score the description by how well another model predicts activations from it (Paulo et al., 2024). The label is therefore a further claim, distinct from all four tiers: that L_i is an accurate description of the property d_i tracks. A latent can be decodable, used and reproducible while its label describes only the most salient slice of its activations. Bolukbasi et al. (2021) documented the corresponding failure for BERT neurons before SAEs existed: a unit can appear to encode one simple concept on the corpus used to inspect it while encoding something broader, and the appearance depends on the narrow slice of language the inspection corpus covers. Section 4 shows that, in the benchmark located here that separates the latent from its label, the label is the weaker of the two.

2.3 What identifiability theory says Tier 4 requires

The question of when a sparse dictionary is unique has a mature literature that predates SAEs. Olshausen and Field (1996) introduced sparse coding as a model of visual cortex, learning a basis from natural images under a sparsity prior. Hillar and Sommer (2011) give conditions under which sparse codes and dictionaries are unique up to the natural symmetries: when data are generated as sparse combinations of a fixed dictionary and the dictionary and supports meet their conditions, any sparsity-constrained procedure that reconstructs the data recovers the generating dictionary. Spielman et al. (2012) prove exact recovery for square dictionaries with sufficiently sparse random coefficients, and Arora et al. (2013) give a polynomial-time algorithm for the overcomplete case when the dictionary is incoherent.

Read as a template for Tier 4, these results have a two-part structure. A uniqueness theorem converts reconstruction success into identification, but only under a generative premise: that the activations at s are (approximately) sparse linear combinations of a fixed set of directions, with sparsity, incoherence and support conditions met. Without the premise there is no theorem, and the broader representation-learning literature says what happens then. Locatello et al. (2019) proved that unsupervised learning of disentangled representations is "fundamentally impossible without inductive biases on both the models and the data", and in more than 12,000 trained models found that well-disentangled solutions could not be picked out without supervision. Identifiability returns when something beyond the observations is supplied: an auxiliary variable on which the latent prior is conditioned (Khemakhem et al., 2020), or sparsity in the mechanisms relating latents to each other and to observed auxiliaries (Lachapelle et al., 2022). Joshi et al. (2025) carry this route into LLM interpretability: their sparse shift autoencoders are

identifiable from paired observations that differ in several unknown concepts but not all, whereas a standard SAE carries no such guarantee.

So a Tier 4 claim about a real language model has a definite shape. It must either (a) compare the dictionary with a known ground truth, or (b) invoke a uniqueness theorem and check its generative premise on M 's activations, or (c) supply auxiliary variation that restores identifiability. Sections 3 to 7 ask which of these has been done.

2.4 A stylised rotation argument

The following is the author's own analytical illustration. It shows how a specific violation of the generative premise produces the pattern of instability that has been reported.

Suppose that, within the activations at s , some property is represented not as a direction but as a two-dimensional circle: the component of the activation in a plane P is $r(\cos t, \sin t)$, where t indexes the property's value. Engels et al. (2024a) found structures of this kind for days of the week and months of the year, and showed by intervention that the circular representation is what models use to compute modular arithmetic over them. Suppose further that t is distributed uniformly and independently of the rest of the activation, so that the distribution of activations is unchanged by the orthogonal map R that rotates P by some angle and leaves its orthogonal complement fixed. An SAE of finite width reconstructs the circle with some number k of latents whose decoder directions lie in or near P . Now apply R to every decoder direction, every encoder direction and the decoder bias of the SAE. The rotated SAE's loss on an input x equals the original SAE's loss on the input $R^T x$, because R preserves norms; and since x and $R^T x$ have the same distribution, the expected reconstruction error and the sparsity of the code are unchanged. Any optimum of the SAE objective therefore comes with a continuum of equally good solutions related by rotation within P , and nothing in the objective selects among them; the optimiser's starting point does. Three consequences follow.

First, the individual latents are not identifiable: which rotation a training run lands on is set by initialisation, so two seeds will produce different latents. Second, the plane P is identifiable within that family: every rotated solution spans the same subspace, so a comparison at the level of subspaces will find agreement where a comparison at the level of latents finds none. Third, the number k is not a property of the representation: a wider SAE can tile the same circle with more latents, each covering a smaller arc, with no width at which the tiling becomes the correct one.

Each consequence has an empirical counterpart. Gerasimov et al. (2026) report that latents which fail to recur across seeds nonetheless concentrate in lower-rank

subspaces that do recur, and in a controlled synthetic model they recover low-rank ground-truth structure at the level of subspaces while it stays non-identifiable at the level of individual latents. Michaud et al. (2025) model how multi-dimensional "feature manifolds" interact with SAE scaling and identify a regime in which manifolds lead an SAE to learn far fewer distinct features than it has latents. Leask et al. (2025) find that larger SAEs contain both finer-grained versions of smaller SAEs' latents and genuinely new ones. When the distribution on the circle is not uniform, the symmetry is broken, but it is broken by the data distribution: the selected rotation then reflects where inputs happen to fall on the circle, not anything about how M processes them. That is a particular instance of the concern Braun et al. (2024) raise in general form, "SAEs may learn more about the structure of the dataset than the computational structure of the network" [sic].

The argument needs one premise that is an empirical claim about M : that some of what M represents at s is multi-dimensional. Engels et al. (2024a) supply direct evidence for some properties; Csordas et al. (2024) give a counterexample to the strong linear representation hypothesis in recurrent networks, where small models store each position's token at a position-specific order of magnitude rather than along a direction; and Hindupur et al. (2025) show across toy, semi-synthetic and naturalistic settings that SAEs which ignore heterogeneity in intrinsic dimensionality fail to recover the concepts that have it.

3. What the standard proxies measure

3.1 Reconstruction, sparsity and the random-weight control

SAE development has been driven by two quantities: how well the dictionary reconstructs activations (often reported as the increase in the model's loss when the SAE's output is spliced in) and how sparse the code is. Both are properties of the pair (activations, procedure). Neither mentions the model's computation, and the field has known this: Gao et al. (2024) introduce further metrics based on recovering hypothesised features, the explainability of activation patterns and the sparsity of downstream effects, and report that all of them generally improve with autoencoder size. Karvonen et al. (2025) assembled SAEBench, eight metrics over more than 200 SAEs spanning eight architectures, and concluded that gains on unsupervised proxy metrics do not reliably carry over to practical performance; Matryoshka SAEs, for example, score slightly worse on the proxies but substantially better on feature disentanglement. Erdogan and Lucic (2025) report the sharper case: on data with rotational symmetries, SAEs that build the symmetry in discover more useful features despite reconstructing worse, so reconstruction quality can run against usefulness.

A particularly informative control in this literature is to train the same SAE on a model whose weights carry no learned computation. Heap et al. (2025) did this across Pythia models from 70M to 6.9B parameters, with several randomisation schemes (weights re-sampled with each matrix's original mean and variance, with or without the trained embeddings, and the step-0 checkpoint), and a control in which each token's input embedding is replaced by fresh Gaussian noise at inference. In many settings, SAEs on the random models produced automated-interpretability scores and reconstruction metrics close to those on the trained model, and much closer to the trained model than to the noise control. The gap was clearer for small models and narrowed for larger ones: auto-interpretability scores on randomised variants were "relatively low for smaller models (e.g., Pythia-70m)" but the difference shrank by Pythia-6.9b. What did separate the conditions was a crude measure of abstractness: latents on randomised models stayed tied to single tokens and did not become more complex with layer depth, whereas latents on the trained model did. Heap et al. draw the appropriate conclusion - high aggregate auto-interpretability scores do not by themselves show that learned, computationally relevant features have been recovered - and recommend randomised baselines as routine.

The result places the most widely reported proxy below Tier 1 as a statement about M . An interpretable latent on a random network is decodable - it reads off a token identity or a simple input pattern - but what it decodes is structure that the data and architecture put into the activations without any learning. A Tier 1 claim about what M has learned to represent therefore needs the random-weight comparison as part of its baseline, and an interpretability score reported without it is evidence about the input distribution as much as about the model.

3.2 Where a ground truth exists, it separates the model from the procedure

The contrast case is the small set of settings where the features a model should represent are known in advance. Karvonen et al. (2024) trained SAEs on language models trained on chess and Othello transcripts, where board-state properties (a knight on a given square, a legal en passant capture, a pin) give a ground-truth feature set, and defined two supervised metrics: coverage (how well the best single latents match each board property) and board reconstruction (how well the board can be recovered from latents). Their Table 1 gives the comparison that Section 3.1 lacked (Figure 1a). On chess, an SAE on the trained model reached 0.48 coverage and 0.85 board reconstruction, while an SAE on a copy of the model with randomly initialised weights reached 0.11 and 0.01. On Othello, the trained model's SAE reached 0.52 and 0.95, against 0.27 and 0.08 for the random copy. A linear probe on the same activations reached 0.98 and 0.98 on chess and 0.99 and 0.99 on Othello.

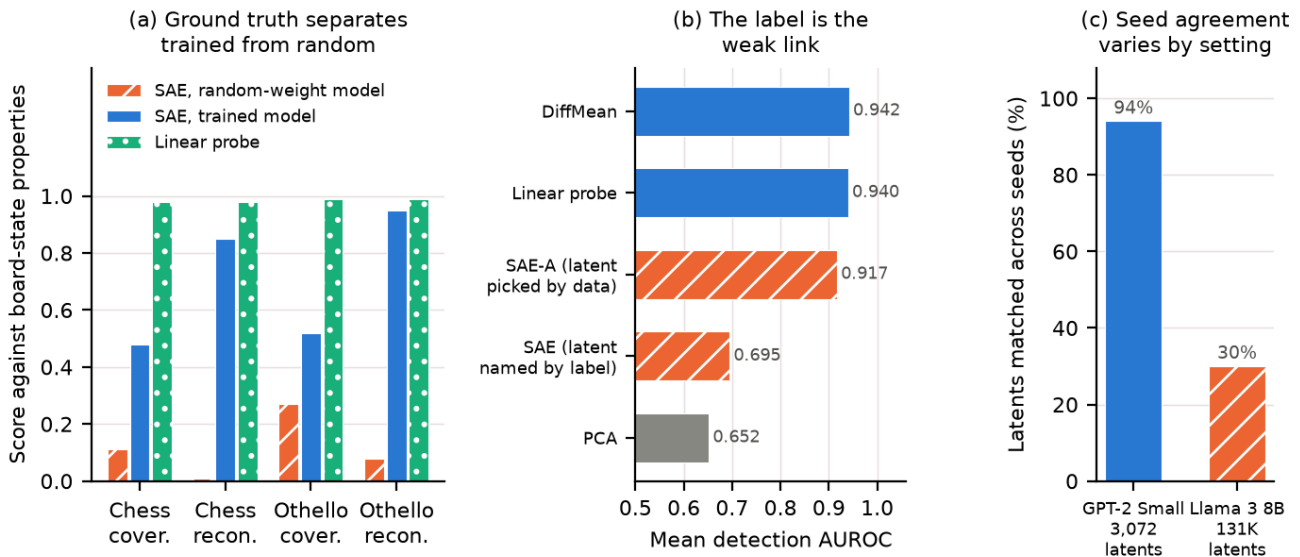


Figure 1. Three published comparisons that separate what a latent is evidence for. (a) Board-game models with a known feature set: best coverage and board-reconstruction scores for SAEs trained on the trained model and on a copy with randomly initialised weights, and for a linear probe (Karvonen et al., 2024, Table 1). (b) Concept detection on Gemma-2 models, mean AUROC averaged over four model-layer settings: the GemmaScope latent whose automated label names the concept (SAE), the latent chosen by detection performance on labelled training data (SAE-A), two supervised baselines and PCA (Wu et al., 2025, Table 1). (c) Share of latents matched across random seeds: 94 percent of latents are reconstruction latents between same-size GPT-2 Small SAEs with 3,072 latents (Leask et al., 2025, appendix, 0.7 cosine threshold) and 30 percent are shared between seeds for a 131K-latent SAE on a Llama 3 8B feedforward layer (Paulo and Belrose, 2025, cosine above 0.7 on both encoder and decoder). The two settings differ in model, width, architecture and matching rule, so panel (c) shows the range reported, not a controlled comparison.

Three readings of these numbers matter. First, with a ground truth, the random-weight control behaves as it should: the procedure applied to an untrained network recovers little, so the trained SAE's scores are evidence about the trained model. The contrast with Heap et al. (2025) is not a contradiction; it shows that the automated proxy and the supervised metric measure different things. Second, the SAE's coverage - a best-single-latent match score averaged over board properties - was around one half, while linear probes on the same activations scored close to 1. The information was present; the dictionary did not carve it at the joints the ground truth defines. Karvonen et al. themselves report per-property detail showing that relational properties fared worst: for the chess model, the best SAE coverage was 0.38 for forks and 0.33 for pins, against probe scores of 0.68 and 0.67. Third, this is a Tier 4 test passed in part, on a model whose world is a game board. It counts as a Tier 4 test because the board state describes the whole world the model is trained on, so the catalogue is meant to cover what there is to represent rather than one task; it is passed only in part, and remains partial as a test, because the catalogue is still chosen by the researchers, which Karvonen et al. note makes their metrics "sensitive to researcher preconceptions". Nothing comparable - a ground truth for the full dictionary - exists for a model trained on natural language.

Other ground-truth settings point the same way. Menon et al. (2024) trained SAEs on models of formal languages (Dyck-2, arithmetic expressions and an English PCFG)

and found that interpretable latents often appear but that results are highly sensitive to the inductive biases of the training pipeline, and that latents correlating with input properties do not always have causal effect on the model's computation. Venhoff et al. (2024) propose SAGE, which derives task-specific ground-truth features by locating task-relevant activations automatically and extends supervised evaluation to Pythia-70M, GPT-2 Small and Gemma-2-2B. Earlier, Makelov et al. (2024) compared unsupervised dictionaries with supervised ones on the indirect object identification (IOI) task in GPT-2 Small; there, SAEs captured interpretable IOI features but were less successful than the supervised dictionaries at controlling the model. These two are the natural-language cases, and they differ in kind from the board games. SAGE describes its targets as "approximate ground truth features" computed from "supervised feature dictionaries" (Venhoff et al., 2024), and both SAGE and Makelov et al. define them for a single task. A ground truth of this kind can test whether an SAE recovers particular variables, but it cannot test whether the dictionary is complete or unique. On the theoretical side, Cui et al. (2025) derive a closed-form solution for SAEs and conclude that they generally fail to recover ground-truth monosemantic features fully unless those features are extremely sparse.

3.3 Compression as a criterion

One proposal makes the choice among decompositions explicit rather than hidden. Ayonrinde et al. (2024) treat an SAE as a lossy compression scheme for communicating explanations of activations and argue from minimum description length for preferring dictionaries that are both accurate and concise, adding a requirement that features be understandable independently. Applied to MNIST, the criterion prefers latents for significant line segments over latents for memorised digits or small fragments. The proposal is useful precisely because it does not pretend to be ontology: it states the preference that selects the decomposition. What it shows, for the argument here, is that once the selection criterion is made explicit, the result is licensed as the best description under that criterion, which is a Tier 3 property of a procedure-family, not a Tier 4 property of the model.

4. Tier 1: decodability, and the weakness of the label

Tier 1 is the claim the record supports best. Cunningham et al. (2023) found SAE latents more interpretable and more monosemantic than directions from alternative decompositions, as judged by automated scoring. Simon and Zou (2024) trained SAEs on the protein language model ESM-2 and report up to 2,548 interpretable latents per layer that correlate strongly with up to 143 known biological concepts, against at most 46 neurons per layer aligned with 15 concepts. Li et al. (2024) find structure in the dictionary itself: parallelogram relations of the kind word embeddings showed, spatially modular "lobes" for related domains such as maths and code, and a power-law spectrum across the whole point cloud. Lan et al. (2024) report that the spaces spanned by SAE features in different language models are highly similar under rotation-invariant comparison, after pairing features by activation correlation. None of this is in dispute here. Each is, at bottom, a claim about structure that can be read from activations by linear means.

The difficulty is what happens when the decodability claim is routed through the latent's label and compared with methods that do not use SAEs. AxBench (Wu et al., 2025) built its concept list from the automated-interpretability labels of GemmaScope latents, generated synthetic data for 500 of those concepts, and asked how well each method detects the concept. The vanilla SAE baseline is the latent whose label names the concept. Its mean AUROC, averaged over four model-layer settings on Gemma-2-2B and Gemma-2-9B, was 0.695 (Figure 1b). Difference-in-means on labelled data reached 0.942 and a linear probe 0.940. The informative row is SAE-A: when the SAE latent is chosen by its detection performance on labelled training data rather than by its label (SAE-A), the mean AUROC rises to 0.917. The dictionary contains a direction that detects the concept nearly as well as supervised methods; the label does not

reliably point to it. Wu et al. note themselves that the automated labels lean towards token-level concepts and miss higher-level abstractions. For steering, the picture is worse: on their overall steering score, prompting averaged 0.894, difference-in-means 0.239 and SAE latents 0.165.

Kantamneni et al. (2025) ran the complementary test, asking whether SAE latents help probing in four hard regimes - scarce data, class imbalance, label noise and covariate shift - where a basis of concept-level latents might be expected to supply a useful inductive bias. SAEs occasionally beat baselines on individual datasets, but no ensemble combining SAEs with baselines consistently beat an ensemble of baselines alone, and applications that initially looked promising (spotting spurious correlations, detecting poor dataset quality, multi-token probes) could be matched by simple non-SAE methods. Enkhbayar (2025), a single-author preprint, applies model-X knockoffs to SAE feature selection for a sentiment task on Pythia-70M and selects 129 of 512 high-activity latents at a target false discovery rate of 0.1, roughly a quarter, as carrying task-relevant signal.

What follows for Tier 1 is narrower than either side's rhetoric. Linear decodability of a large number of properties at a site is well established, and SAEs are one efficient, unsupervised way to enumerate candidate directions. The specific claim "latent i encodes the property its label names" is weaker than it looks, because the label is generated from the most salient contexts and evaluated by a proxy that, by Section 3.1, does not separate learned structure from input structure. The AxBench SAE-A row suggests a practical repair: treat the label as a hypothesis and select latents against labelled held-out data, at which point the SAE is being used as a candidate generator for a supervised test.

5. Tier 2: use, and its dependence on the family

5.1 Evidence that latents matter causally

Tier 2 evidence has grown fast and is genuinely mixed. On the positive side, Cunningham et al. (2023) located the latents responsible for counterfactual behaviour on IOI more finely than earlier decompositions. Marks et al. (2024) built sparse feature circuits - causally implicated subgraphs of SAE latents - and used them for SHIFT, improving a classifier's generalisation by ablating latents a human judged irrelevant to the task. Cho et al. (2026) ran a large direct test of necessity: they searched 3.9 million SAE latents across six models and three SAE families for single-token latents (which fire on one vocabulary item, so the expected effect is known in advance). These are a small fraction of the total - about 26,600 entered their causal analysis - and they tested them by zero-ablation. Deleting such a latent lowered the model's logit for its token in 178 of 208 layer

conditions after correction for multiple comparisons. Arad et al. (2025) separate latents that mainly reflect patterns in the input from latents whose activation has a human-readable effect on the output, find that high input and output scores rarely co-occur in one latent, and report two- to three-fold improvements in steering when low-output-score latents are filtered out, enough to make SAE steering competitive with supervised methods.

5.2 Evidence that the causal role belongs to the procedure

The same studies carry the qualification. Cho et al. (2026) found that causal role depended on the SAE family more than on scale within a family. GemmaScope and BatchTopK latents, compared on the same Gemma base models, were causally anchored, while LlamaScope latents, measured on Llama-3.1-8B and DeepSeek-R1-8B, were locally redundant: after ablation the token returned to within twice its pre-ablation rank 96 to 98 percent of the time. Changing only the activation function reversed the sign of the difference, leaving the training recipe as the remaining candidate. A causal claim made with one family therefore cannot be assumed to transfer to another, even for single-token latents, the simplest case available. Arad et al.'s separation of input and output latents suggests the same thing from a different side: since high input and output scores rarely co-occur, the latent chosen because its activating contexts match a concept will often not be the one whose intervention produces that concept.

Hoang et al. (2026) study the geometry of what an intervention does rather than where a latent sits. Removing the same active latent across many contexts and examining the cloud of resulting logit changes, they find consistent one-dimensional effects rare across SAE variants; latents tied to static information such as factual attributes more often have structured, low-dimensional (though usually multi-directional) effects, while latents associated with context-dependent operations mostly have diffuse ones. A latent can thus be interpretable and causally relevant without providing a stable direction to steer along. Farrell et al. (2024), applying SAE latents to unlearning on the biology subset of WMDP, found that zero-ablating latents was ineffective, that negative scaling was needed, and that intervening on several latents together produced side effects similar to or larger than an existing fine-tuning-based method. Muhamed et al. (2025) later report that a dynamic, classifier-gated use of SAE latents beats leading unlearning methods; the contrast is between ways of using latents, not evidence that the latents themselves changed.

The causal-variable benchmarks deliver the cleanest comparisons. On RAVEL (Huang et al., 2024), which asks whether a method can isolate, for example, a city's country from its continent in Llama2-7B, supervised multi-task distributed alignment search performed best. Chaudhary and Geiger (2024) applied RAVEL to four

open-source SAEs on GPT-2 Small, learning masks over latents to change a city's country without changing its continent, and found that the SAEs struggled to reach the neuron baseline and none came near the distributed alignment search skyline. MIB (Mueller et al., 2025) generalised this into a causal-variable localisation track across tasks and models and reported that SAE features were no better than neurons; in their words, "DBM on SAE features is the worst performing method other than the full vector baselines". Causal abstraction (Geiger et al., 2023) supplies the theory under which these comparisons are meaningful: an SAE is one way of choosing a featurisation of a hidden vector, and the benchmarks ask whether the chosen features align with a high-level causal variable.

5.3 Interventions can manufacture their own evidence

A last caution bears on Tier 2 directly. Makelov et al. (2023) show that a subspace intervention can make the model behave as though a feature's value had changed by activating a dormant parallel pathway through a subspace that is causally disconnected from the outputs in normal operation, so that manipulating behaviour and attributing it to a subspace come apart. An intervention on an SAE latent is a subspace intervention. Its success licenses the claim that the direction can be used to move the model; it does not by itself license the claim that the model uses that direction when left alone. Tier 2 evidence should therefore be read with the intervention type attached, and claims of the form "the model uses latent i for X " need interventions within the activation distribution the model actually produces, which is what mean-ablation and resample-ablation designs try to approximate.

Two design choices change what Tier 2 means. Braun et al. (2024) train SAEs end to end, minimising the divergence between the original model's output distribution and the output with SAE activations inserted, rather than activation reconstruction; this targets functional importance directly and gives a Pareto improvement on explained performance, total features and active features per input. Dunefsky et al. (2024) introduce transcoders, which approximate an MLP's input-output map with a wider sparse layer, and Paulo et al. (2025) report that transcoder latents are significantly more interpretable than SAE latents trained on the same model and data. Both are improvements. Both also show that what a "feature" is depends on the objective: reconstruct the activation, reconstruct the downstream effect, or reconstruct a layer's function. These are three different decompositions of one model, each defensible, and nothing in the evidence reviewed here privileges one of them as the model's own.

6. Tier 3: reproducibility, at which grain

6.1 Latents do not reliably recur

The seed experiment is the cheapest Tier 3 test, and its results disagree by setting. Paulo and Belrose (2025) matched latents between SAEs that differed only in initialisation, counting a latent as shared when its best match had cosine similarity above 0.7 on both encoder and decoder. For a 131K-latent SAE on a feedforward layer of Llama 3 8B, about 30 percent of latents were shared. The pattern held across layers of three models, two datasets and several architectures; smaller models and smaller SAEs shared more; and ReLU SAEs with an L1 penalty were more stable than TopK SAEs at matched sparsity. The unmatched "orphan" latents were often interpretable, so a single training run misses interpretable directions that another run finds. Leask et al. (2025), on GPT-2 Small SAEs with 3,072 latents, report that 94 percent of latents in one seed were reconstruction latents (matched at a 0.7 cosine threshold) for an SAE of the same size from another seed (Figure 1c). The two numbers do not contradict each other; they come from models about sixty times apart in parameter count, dictionaries about forty times apart in width, and different training setups and matching rules. What they establish together is that latent-level reproducibility is a property of the setting and not of SAEs in general; Paulo and Belrose's own observation that smaller models and smaller SAEs share more suggests that it degrades in the direction the field is scaling.

Paulo and Belrose (2025) also record a detail that bears on how Tier 3 evidence is read. A positive result on feature matching across adjacent layers of the Gemma 2 models was, they report from personal communication with the team that trained GemmaScope, obtained on SAEs that all shared one random seed, and they judge that result likely to depend on that choice. Shared initialisation can make dictionaries agree for reasons unrelated to the activations, so cross-dictionary agreement is evidence of reproducibility only when the dictionaries were trained independently.

Instability is not confined to language. Fel et al. (2025) report severe instability in SAEs for large vision models, where identical models trained on similar datasets can produce sharply different dictionaries, and propose archetypal SAEs that constrain dictionary atoms to the convex hull of the data. The constraint markedly improves stability, a mildly relaxed variant matches state-of-the-art reconstruction, and both are assessed on new benchmarks of plausibility (recovering known classification directions) and identifiability (separating synthetic concept mixtures).

6.2 Reproducibility can be engineered, and what that buys

Song et al. (2025) argue that consistency across runs should be a first-class target and propose the pairwise dictionary mean correlation coefficient (PW-MCC) to measure it. In a synthetic model organism with known features, PW-MCC tracks ground-truth recovery; on LLM activations, TopK SAEs with appropriate architectural choices reach 0.80, and higher consistency correlates strongly with the semantic similarity of the latents' automated explanations. Nelson et al. (2026) show theoretically that standard SAEs are expected to be unstable, identify the model properties responsible, and propose identifiable SAE variants of the TopK SAE with lower reconstruction error and better stability. They connect the improvement to classical dictionary learning and report that dictionaries learned in practice satisfy an approximate restricted isometry condition, which makes the corresponding sparse codes near-identifiable. Korznikov et al. (2025) penalise pairwise cosine similarity between decoder directions and report 9 percent more distinct features, 65 percent less absorption and 15 percent less composition, a 6 percent gain on spurious-correlation removal and on-par results on their other downstream tasks.

These results are real progress at Tier 3, and their scope needs stating exactly. Consistency is agreement between runs of a procedure; a procedure can be made to agree with itself by adding constraints that pin the solution, and the pinned solution is then reproducible because of the constraint. That is the right outcome if the constraint matches how the activations were generated, and the wrong one if it does not. Song et al.'s synthetic validation is the right kind of step for closing this gap - in a model organism where the generative process is known, consistency predicts recovery - and the open question is whether the same holds for LLM activations, where the generative process is the thing in question. Nelson et al.'s near-identifiability result concerns the codes given a dictionary that satisfies an isometry condition; it does not by itself make the dictionary unique, and it inherits the generative premise of Section 2.3.

6.3 Subspaces recur where latents do not

Gerasimov et al. (2026) supply the finding that reorganises this section. Estimating, for each latent, the probability that a similar one reappears in an independently trained SAE, they separate stable from unstable latents across seeds, models, layers, widths and SAE variants. Stable latents carry most of the reconstruction-relevant and prediction-relevant signal; masking many unstable latents changes explained variance only slightly, while masking far fewer stable ones produces a larger drop. Unstable latents fire less often (mean activation frequency 0.18 percent against 0.44 percent for stable ones) and their activations and explanations are dominated by low-frequency surface

triggers. Geometrically, the unstable latents are individually non-reproducible but jointly occupy lower-rank subspaces that do reproduce. This is the signature that the rotation argument in Section 2.4 predicts: an identifiable span with a basis chosen by initialisation.

If that reading is right, the debate between realists and deflationists has been conducted at the wrong grain. The realist claim fails for individual latents in the unstable fraction, and may well hold for subspaces. The deflationist claim that a dictionary is one useful basis among many is correct within those subspaces, and may be too strong for the stable fraction. A Tier 3 claim should therefore state its grain: "this latent recurs" and "this subspace recurs" are different claims with different evidence, and the second is the one the current record supports more readily.

6.4 Across models and across fine-tuning

Cross-model agreement is a stronger form of Tier 3, and it has been reported: Lan et al. (2024) find high similarity between SAE feature spaces across different LLMs once features are paired by activation correlation, and Thasarathan et al. (2025) train a single universal SAE that reconstructs the activations of several vision models from any one of them, recovering shared concepts from colours and textures to parts and objects. Both results compare spaces after an alignment step, which is the subspace grain again; neither shows that the individual latents of independently trained dictionaries correspond. Model diffing exposes how much the procedure can contribute to such comparisons. Minder et al. (2025) show that crosscoders, which learn a shared dictionary for a base model and its chat-tuned version, inherit artifacts from their L1 penalty that make concepts present in both models appear unique to the fine-tuned one, and that the standard crosscoder suffered heavily from this when comparing Gemma 2 2B base and chat. A claim that fine-tuning "introduced" a latent is a Tier 3 claim about two models, and here the sparsity objective manufactured part of the evidence for it.

7. Tier 4: canonicity, and the premise nobody has checked

7.1 Not complete, not atomic

Leask et al. (2025) test Tier 4 directly. SAE stitching inserts latents from a larger SAE into a smaller one: some larger-SAE latents substitute for smaller-SAE latents with similar behaviour, while others are novel, improving reconstruction when added and so carrying information the smaller dictionary missed. The existence of novel latents shows the smaller dictionaries were incomplete. Meta-SAEs, trained on the decoder directions of another SAE, show that latents in larger SAEs often decompose into combinations of latents from smaller ones - their example is a latent for "Einstein" splitting into "scientist", "Germany" and "famous person" - so the larger latents are

not atomic. A meta-SAE in their setup explained 55.47 percent of the variance of the decoder directions it was trained on. Both properties that Tier 4 asks for, completeness and atomicity, fail as a function of width, and there is no width at which the evidence says they begin to hold.

7.2 The sparsity objective rewrites hierarchy and correlation

Chanin et al. (2024) identify feature absorption: when features form a hierarchy, a latent that appears to track a general property stops firing on inputs where a more specific child latent fires, because the sparsity penalty rewards using one latent where two would be faithful. They show this is a consequence of optimising for sparsity whenever features are hierarchical, validate a metric for it on hundreds of LLM SAEs, and conclude that varying width or sparsity is not enough to remove it. Chanin et al. (2025) identify the converse failure for narrow SAEs: when the dictionary has fewer latents than there are underlying features and those features are correlated, the reconstruction loss merges parts of correlated features into one latent. They call this feature hedging and judge both of its conditions almost certainly met in LLM SAEs. Their conclusion is that width is not a neutral hyperparameter. Matryoshka SAEs (Bussmann et al., 2025) address absorption by training nested dictionaries that must each reconstruct alone, so that general concepts survive at small widths. Each fix is a choice about which hierarchy-handling the dictionary should embody.

Absorption bears on the most demanding use of a dictionary: reading the absence of activation as the absence of a property. If a parent latent fails to fire precisely where a child latent absorbs it, then "this latent did not fire" does not imply "the model did not represent this property here". A negative claim needs completeness, which Section 7.1 shows is not available, and non-absorption, which this section shows is not guaranteed.

7.3 Reconstruction error is not all missing features

Engels et al. (2024b) study what SAEs leave unexplained. About half of the error vector, and more than 90 percent of its norm, can be predicted linearly from the original activation. The part that cannot - the nonlinear error - contains fewer not-yet-learned features, yields worse SAEs when SAEs are trained on it, and accounts for a proportional share of the increase in loss when SAE reconstructions are inserted into the model. Larger SAEs mostly struggle on the same contexts as smaller ones. On the encoding side, O'Neill et al. (2025) use compressed-sensing theory to prove that the simple linear-then-nonlinear encoder of an SAE cannot perform accurate sparse inference even in cases where the sparse code is recoverable, and show that more expressive encoders improve both inference and interpretability. Taken together, these results suggest that even if a

canonical dictionary existed, a standard SAE would neither reconstruct the activations fully in its terms nor reliably assign the correct codes.

7.4 The generative premise, item by item

Section 2.3 reduced Tier 4 to three routes: a ground truth, a uniqueness theorem with its premise checked, or auxiliary variation that restores identifiability. Route (a) has been taken for board-game and formal-language models (Section 3.2) and for synthetic model organisms (Song et al., 2025; Gerasimov et al., 2026); it has not been taken for a model trained on natural language, because there is no ground truth for the full dictionary to take it with. The task-specific, approximate ground truths of SAGE (Venhoff et al., 2024) and Makelov et al. (2024) test the recovery of particular variables, not completeness or uniqueness. Route (c) has been opened by sparse shift autoencoders (Joshi et al., 2025), which recover concepts from paired observations, but it yields identifiability for the concepts that vary across pairs, not a complete dictionary. That leaves route (b), and the premise it would need to check is that activations at a site are sparse linear combinations of a fixed set of directions satisfying the uniqueness conditions. The published record bears on each part of that premise.

Linearity. Park et al. (2024) formalise the linear representation hypothesis in counterfactual terms, connect its two forms to probing and steering, and report experiments on LLaMA-2 that exhibit linear representations of concepts, which supports the hypothesis for the concepts tested. Engels et al. (2024a) find irreducibly multi-dimensional structure for others; Csordas et al. (2024) find non-linear, magnitude-based representations in small recurrent networks; and Hindupur et al. (2025) argue that each SAE architecture builds in an assumption about intrinsic dimensionality and separability and recovers only the concepts that match it, concluding that an SAE does not only reveal concepts but "determines what can be seen at all".

Sparsity and independence of supports. Absorption exists because features are hierarchical (Chanin et al., 2024), and hedging because they are correlated (Chanin et al., 2025); both depart from the random, well-spread supports that the recovery results of Section 2.3 assume. Sun et al. (2025) complicate the picture from the other side: latents that fire on a large fraction of inputs, long suspected to be training artifacts, behave as a property of the residual space (when their subspace is ablated, retrained SAEs stop producing new dense latents), tend to come in antipodal pairs that reconstruct particular residual-stream directions, and fall into interpretable classes such as position tracking and entropy regulation. Some of what the model represents at a site is dense, not sparse.

Fixed dictionary. The nonlinear residual (Engels et al., 2024b) and the manifold regime (Michaud et al., 2025) both suggest that some of the variance at a site is not

well described by any finite set of directions.

None of this refutes the premise. Each item shows that it fails for some portion of what models represent, and none of the work estimates how large that portion is. That is the precise sense in which no real-model result has checked the premise: the parts of it that have been tested were tested on specific concepts chosen because they were tractable, and no study has estimated what fraction of a real model's activation variance at a site is generated by a code that meets the conditions a uniqueness theorem needs. A Tier 4 claim about an LLM is currently a claim made without that estimate.

8. The strongest case for the programme

The case for SAEs, stated as strongly as the evidence allows, does not need Tier 4, and several of its best results do not reach for it.

The first argument is discovery. Peng et al. (2025) argue that SAEs may be weak tools for acting on concepts one already knows but strong tools for finding concepts one does not, and that this distinction sorts the negative results from the positive ones. HypotheSAEs (Movva et al., 2025) is the cleanest example: an SAE trained on text embeddings, latents selected for predicting a target such as headline clicks, and an LLM-generated interpretation of each selected latent offered as a hypothesis. Against baselines, it recovered reference hypotheses better on synthetic data (at least 0.06 higher F1) and yielded about twice as many significant findings on real data, at one to two orders of magnitude less compute than LLM-based alternatives. The important feature of this design is where validation happens. In their evaluation, each hypothesis is annotated on a held-out set by a separate language model and counted only if it remains significant, after Bonferroni correction, in a regression on the target - a test that does not involve the SAE. The SAE is a candidate generator, and what is claimed at the end is a fact about the dataset, licensed by the test, not a fact about the embedding model.

The second argument is that the instability is partly cosmetic. Gerasimov et al. (2026) find that stable latents carry most of the reconstruction and prediction signal, and that pooling unique latents across seeds yields more stable SAEs while keeping explained variance, in the setting they test. If the unstable fraction is mostly low-frequency surface triggers inside reproducible subspaces, then the part of the dictionary that matters for most purposes is the part that reproduces. This is a substantial defence of Tier 3 at the subspace grain and of Tier 3 at the latent grain for the stable fraction.

The third argument is utility in audits. In a blind auditing game, three of four teams uncovered a hidden objective deliberately trained into a model, using techniques that included SAE-based interpretability alongside behavioural attacks and training-data analysis (Marks et al., 2025). Simon and Zou (2024) report that

latents in a protein language model can fill in missing database annotations and steer sequence generation. These are successes of the method as a search tool whose outputs are then checked by other means.

The fourth argument is that canonicity is the wrong target, and that the programme never needed it. Sharkey et al. (2025), in a broad review of open problems, and Bereska and Gavves (2024), in a review of mechanistic interpretability for safety, both present the field's methods as needing conceptual clarification and not only better engineering. Ayonrinde and Jaburi (2025) go further, proposing that mechanistic interpretability produces model-level, ontic, causal-mechanistic and falsifiable explanations, and naming a principle of explanatory optimism as the conjecture the field depends on. On their framing, the realist commitment is made openly, as a hypothesis about what models contain, which is the honest way to make it.

This paper accepts all four arguments and draws one conclusion from them: the uses that the evidence supports are uses in which the SAE's output is validated by a test the SAE does not supply - a held-out statistical test, an intervention with comparison directions, a behavioural check in an audit. The uses in trouble are those in which the dictionary's structure is itself the evidence: reading a latent's label as what the model computes, reading a missing latent as a missing representation, reading the number of latents as the number of things the model knows, and reading a latent unique to one model as a difference between models.

9. A use-indexed licensing procedure

9.1 The evidence, consolidated

Table 1 collects the measured values the argument above rests on, one row per result, with the tier each bears on. Three patterns are visible in it. In the table, the values that separate a trained model from an untrained one come from settings with a ground truth or a supervised criterion (rows 1 and 2 against row 4). Where an SAE route is compared with a supervised non-SAE route to the same end, the supervised route wins in rows 3, 5 and 6. Selecting the SAE latent with supervision narrows the gap for detection (row 5) but not for steering, where AxBench's data-selected latent scored 0.157 against 0.165 for the label-named one (row 6), nor for relational board properties, where coverage already takes the best-matching latent per property (row 3); only selection by output score makes SAE steering competitive (row 9). The comparison with raw neurons favours SAEs on interpretability counts (row 17), though not on causal-variable localisation (Mueller et al., 2025; Chaudhary and Geiger, 2024). And the reproducibility values span the range from 30 to 94 percent depending on setting (rows 10 and 11), with the most stable reported grain being a subspace (Gerasimov et al., 2026; Section 6.3).

9.2 The procedure

Algorithm 1 states the decision structure this paper recommends for any claim made with an SAE latent. Its organising rule is that the evidence required is set by the use, and that a claim is stated at the highest tier its evidence reaches and no higher.

Table 2 gives the tier each common use requires and the check that must come from outside the SAE.

9.3 Three published cases, read through the procedure

Sparse feature circuits (Marks et al., 2024) make a Tier 2 claim - these latents are causally implicated in this behaviour - with interventions, and use it for SHIFT, a Tier 2 use whose success is validated by the downstream classifier's generalisation. Algorithm 1 licenses the use as reported. It would require two additions before the circuit is read as the mechanism: replication in a second SAE family (step 4), given Cho et al.'s finding that causal roles change across families, and a statement of grain (step 5).

The hidden-objective audit (Marks et al., 2025) used SAEs as one search tool among several and confirmed findings by other means. It is a Tier 1 to 2 use with external validation, and the procedure licenses its conclusion. What the procedure does not license is the converse inference, which the audit did not make: that an audit of a different model in which no suspicious latent appears has shown the objective to be absent. That is a Tier 4 claim (Table 2), and absorption, incompleteness and the nonlinear residual (Section 7) all stand between "no latent fired" and "nothing is there".

Crosscoder-based model diffing (Minder et al., 2025) is a Tier 3 claim across two models. The procedure's step 5 asks for independent training and artifact checks, and Minder et al.'s latent scaling supplies the artifact half; without it, the claim that fine-tuning introduced a concept is partly produced by the L1 objective.

10. What would settle the open part

The open part is Tier 4 for a model trained on natural language, and the grain question at Tier 3. Five studies would move it. None needs new theory, and the first three are within the compute budget of groups already publishing SAE suites.

S1. Random-weight controls as standard reporting.

Every reported Tier 1 result - an interpretability score, a count of interpretable latents, a concept-correlation figure such as the 2,548 latents of Simon and Zou (2024) - should be accompanied by the same pipeline applied to a copy of the model with re-sampled weights, as Heap et al. (2025) and Karvonen et al. (2024) did. The prediction, if the deflationary reading is right for automated metrics, is that the trained-random gap will be small for token-level latents and large for latents that require composition; if the gap is large everywhere, the concern of Section 3.1 is overstated.

Table 1. Measured values the argument rests on, one row per cited result. "Tier" is the claim of Section 2.1 that the value bears on most directly; "label" is the label-fidelity claim of Section 2.2. All values are as reported by the source on the row and are consolidated here by the author.

#	Source	Setting	Measure	Value	Tier
1	Karvonen et al. (2024), Table 1	Chess LM	SAE coverage / board recon., trained vs random weights	0.48 / 0.85 vs 0.11 / 0.01	4 (partial)
2	Karvonen et al. (2024), Table 1	Othello LM	same	0.52 / 0.95 vs 0.27 / 0.08	4 (partial)
3	Karvonen et al. (2024), Table 4	Chess LM	best SAE coverage vs probe, fork / pin	0.38 / 0.33 vs 0.68 / 0.67	1, 4
4	Heap et al. (2025)	Pythia 70M-6.9B	auto-interp score, trained vs random weights	similar in many settings; gap narrows with size	1 (baseline)
5	Wu et al. (2025), Table 1	Gemma-2 2B/9B, 500 concepts	detection AUROC: label-named latent / data-selected latent / DiffMean	0.695 / 0.917 / 0.942	label, 1
6	Wu et al. (2025), Table 2	same	steering score: SAE / SAE-A / DiffMean / prompting	0.165 / 0.157 / 0.239 / 0.894	2
7	Cho et al. (2026)	6 models, 3 SAE families	layer conditions where ablation lowers own-token logit	178 of 208	2
8	Cho et al. (2026)	LlamaScope latents	token back within 2x rank after ablation	96-98%	2 (family)
9	Arad et al. (2025)	SAE steering	gain after filtering by output score	2-3x	2
10	Paulo and Belrose (2025)	Llama 3 8B FFN, 131K latents	latents shared across seeds	30%	3
11	Leask et al. (2025)	GPT-2 Small, 3,072 latents	same-size reconstruction latents across seeds	94%	3
12	Song et al. (2025)	LLM activations, TopK	pairwise dictionary MCC	0.80	3
13	Gerasimov et al. (2026)	several models and SAEs	firing frequency, unstable vs stable latents	0.18% vs 0.44%	3
14	Korznikov et al. (2025)	OrtSAE vs standard	change in absorption / composition	-65% / -15%	3, 4
15	Leask et al. (2025)	meta-SAE on decoder directions	variance explained	55.47%	4 (atomicity)
16	Engels et al. (2024b)	SAE error vectors	linearly predictable share of error / of its norm	about half / over 90%	4 (completeness)
17	Simon and Zou (2024)	ESM-2 per layer	interpretable latents vs concept-aligned neurons	up to 2,548 vs up to 46	1
18	Enkhbayar (2025)	Pythia-70M, sentiment	latents selected at FDR 0.1	129 of 512	1
19	Movva et al. (2025)	HypotheSAEs	F1 gain (synthetic); significant findings (real)	at least +0.06; about 2x	1 (data)

Algorithm 1. Licensing a claim made with an SAE latent (or set of latents)

Input: claim C about model M at site s; dictionary D from procedure P; use U step 1 classify C: T1 decodable T2 used T3 reproducible T4 canonical (absence claims, concept counts and "the model's features" are T4) step 2 look up the tier U requires (Table 2); if tier(C) > tier(U) then restate C at tier(U) step 3 T1: test the label on labelled held-out data against a probe or DiffMean baseline; repeat on a random-weight copy of M; report both margins step 4 T2: intervene with a stated type (zero / mean / resample / steer); compare with matched random and supervised directions; replicate in a second SAE family before generalising step 5 T3: retrain P with new seeds and widths, no shared initialisation; report per-latent recurrence AND subspace recurrence; state C at the grain that recurs step 6 T4: require one of (a) agreement with a known ground truth, (b) a uniqueness theorem whose generative premise is checked at s, (c) auxiliary variation that restores identifiability; if none holds then C is not licensed at T4: report the T3 claim step 7 output C with its tier, grain, intervention type and controls
--

S2. Planted-structure language models. Train small language models on text whose generating process includes latent variables of known form - one-dimensional, circular, hierarchical, and correlated - embedded in otherwise natural-looking text, and measure how SAEs of varied width, architecture and

seed recover each kind. Board games (Karvonen et al., 2024) and formal languages (Menon et al., 2024) are the precedents; the step needed is a ground truth that coexists with the statistical texture of natural language. The study would turn the four items of Section 7.4 from case-by-case observations into measured recovery rates

Table 2. Uses of SAE latents, the tier each requires, the validation that has to come from outside the dictionary, and the state of the published evidence. The assignments are this paper's; the status column summarises Sections 3 to 8.

Use	Tier needed	Check outside the SAE	Status in the record
Hypotheses about a dataset	1 (about data)	held-out statistical test	supported (Movva et al., 2025)
Probing, concept detection	1	labelled held-out data, non-SAE baseline	no consistent gain (Kantamneni et al., 2025; Wu et al., 2025)
Steering, editing	2 (output side)	comparison directions, side effects	mixed; better with output selection (Arad et al., 2025)
Unlearning	2 plus robustness	relearning and side-effect tests	mixed (Farrell et al., 2024; Muhamed et al., 2025)
Circuit explanation	2 and 3	in-distribution ablation, second family	family-dependent (Cho et al., 2026)
Model diffing	3 across models	independent dictionaries, artifact checks	L1 artifacts (Minder et al., 2025)
Audit: finding what is present	1-2, as search	behavioural confirmation	supported in one game (Marks et al., 2025)
Audit: certifying absence	4	none available	unsupported
Counting a model's concepts	4	none available	unsupported

per structure type.

S3. Grain-explicit replication. For a public SAE suite, report for every latent the probability that it recurs across independent seeds (the quantity Gerasimov et al., 2026, estimate) and, for the unstable fraction, the principal angles between the subspaces it occupies across seeds, widths and families. The output would be a map of which claims can be stated per latent and which only per subspace, and it would test the rotation argument of Section 2.4 directly: if that argument is right, the unstable latents with reproducible spans should be enriched for properties with known multi-dimensional structure.

S4. An estimate of the premise share. For a given site, decompose the activation variance into the part reconstructed by stable, recurrent, sparse latents; the part reconstructed only by seed-dependent latents; the dense part (Sun et al., 2025); and the nonlinear residual (Engels et al., 2024b). The first share is an upper bound on how much of the site a Tier 4 claim could ever cover. No published study reports it.

S5. A miss-rate study for absence claims. Fine-tune a model to carry a known property - a planted objective, a backdoor, a factual association - at several strengths, then run SAE-based audits without knowledge of the property, across SAE families and widths, and record how often the property is missed. The hidden-objective game of Marks et al. (2025) is the template for the positive direction; the negative direction, the rate at which SAE audits fail to find what is there, is the number any claim of absence would need and the one the record lacks.

11. What this paper does not show

This paper consolidates and analyses published measurements, and its claims are about what those measurements license.

The four tiers are stipulations. Other decompositions are possible; Saphra and Wiegrefe (2024) distinguish

several senses of "mechanistic", and the philosophical literature that Williams et al. (2025) and Vilas et al. (2024) bring to interpretability could carve the claims differently. The tiers are offered because they map onto distinct evidence requirements, not because they are the only map.

The rotation argument of Section 2.4 is stylised. It assumes an exactly rotation-symmetric planar component, which real activations need not have, and it explains one source of instability among several. It is consistent with the subspace result of Gerasimov et al. (2026) and the manifold model of Michaud et al. (2025); it has not been tested against them.

The evidence base is recent and concentrated. Many of the results that carry the argument are 2025 and 2026 preprints (Cho et al., 2026; Gerasimov et al., 2026; Hoang et al., 2026; Michaud et al., 2025; Enkhbayar, 2025), several come from a small number of research groups, and some of the most negative and most positive results come from the same groups. The record could change quickly. In particular, if identifiable SAE variants (Nelson et al., 2026) or consistency-optimised training (Song et al., 2025) were shown to recover a planted ground truth in a natural-language setting, the Tier 4 verdict here would need revision, and Section 10 is written so that such a result would count.

Full texts were read for the sources whose numbers carry the most weight - Paulo and Belrose (2025), Leask et al. (2025), Heap et al. (2025), Karvonen et al. (2024), Wu et al. (2025), Mueller et al. (2025), Gerasimov et al. (2026), Cho et al. (2026) and Movva et al. (2025). Everything else, including Song et al. (2025), Nelson et al. (2026), Peng et al. (2025), Fel et al. (2025), Joshi et al. (2025) and the identifiability results of Section 2.3, is carried at abstract level, and the claims attached to those sources are limited to what their abstracts state. The comparisons in Figure 1c and rows 10 to 12 of Table 1 come from different settings and matching rules and are presented as a range, not as a controlled contrast.

Finally, this paper does not show that language models lack canonical units. It shows that no published test on a model trained on natural language was built so that it could confirm them if they exist, and that the tests which could be specified.

12. Conclusion

An SAE latent is evidence for something, and the question has been which thing. Read against four claims, the published record supports decodability broadly, with the label as the weak link; it supports causal use unevenly and in a way indexed to the SAE family; it supports reproducibility more readily at the grain of subspaces than of individual latents; and it supports canonicity on no model trained on natural language, because the premise a uniqueness theorem would need has been tested only piecemeal and its coverage never estimated. The uses that hold up are those whose conclusions are checked by something outside the dictionary. The uses that do not are those in which the dictionary's own structure - its labels, its gaps, its count, its differences between models - is taken as the finding. A dictionary is a description of activations under a stated procedure. Whether it is also an ontology of the model is an empirical question with a known form, and it remains open.

References

- Arad, D., Mueller, A., & Belinkov, Y. (2025). SAEs Are Good for Steering – If You Select the Right Features. arXiv:2505.20063. doi:10.48550/arXiv.2505.20063. [PREPRINT]
- Arora, S., Ge, R., & Moitra, A. (2013). New Algorithms for Learning Incoherent and Overcomplete Dictionaries. arXiv:1308.6273. doi:10.48550/arXiv.1308.6273.
- Ayonrinde, K., & Jaburi, L. (2025). A Mathematical Philosophy of Explanations in Mechanistic Interpretability – The Strange Science Part I.i. arXiv:2505.00808. doi:10.48550/arXiv.2505.00808. [PREPRINT]
- Ayonrinde, K., Pearce, M. T., & Sharkey, L. (2024). Interpretability as Compression: Reconsidering SAE Explanations of Neural Activations with MDL-SAEs. arXiv:2410.11179. doi:10.48550/arXiv.2410.11179. [PREPRINT]
- Bereska, L., & Gavves, E. (2024). Mechanistic Interpretability for AI Safety – A Review. Transactions on Machine Learning Research. arXiv:2404.14082. doi:10.48550/arXiv.2404.14082.
- Bolukbasi, T., Pearce, A., Yuan, A., Coenen, A., Reif, E., Viegas, F., & Wattenberg, M. (2021). An Interpretability Illusion for BERT. arXiv:2104.07143. doi:10.48550/arXiv.2104.07143. [PREPRINT]
- Braun, D., Taylor, J., Goldowsky-Dill, N., & Sharkey, L. (2024). Identifying Functionally Important Features with End-to-End Sparse Dictionary Learning. arXiv:2405.12241. doi:10.48550/arXiv.2405.12241.
- Bussmann, B., Leask, P., & Nanda, N. (2024). BatchTopK Sparse Autoencoders. arXiv:2412.06410. doi:10.48550/arXiv.2412.06410. [PREPRINT]
- Bussmann, B., Nabeshima, N., Karvonen, A., & Nanda, N. (2025). Learning Multi-Level Features with Matryoshka Sparse Autoencoders. arXiv:2503.17547. doi:10.48550/arXiv.2503.17547. [PREPRINT]
- Chanin, D., Dulka, T., & Garriga-Alonso, A. (2025). Feature Hedging: Correlated Features Break Narrow Sparse Autoencoders. arXiv:2505.11756. doi:10.48550/arXiv.2505.11756. [PREPRINT]
- Chanin, D., Wilken-Smith, J., Dulka, T., Bhatnagar, H., Golechha, S., & Bloom, J. (2024). A is for Absorption: Studying Feature Splitting and Absorption in Sparse Autoencoders. NeurIPS 2025. arXiv:2409.14507. doi:10.48550/arXiv.2409.14507.
- Chaudhary, M., & Geiger, A. (2024). Evaluating Open-Source Sparse Autoencoders on Disentangling Factual Knowledge in GPT-2 Small. arXiv:2409.04478. doi:10.48550/arXiv.2409.04478. [PREPRINT]
- Cho, S., Wu, Z., Da Costa, K., Kalra, R., Wicaksono, I., & Koshiyama, A. (2026). Are Single-Token Sparse Autoencoder Features Causally Necessary? Layer-Depth and SAE-Family Effects. arXiv:2607.20596. doi:10.48550/arXiv.2607.20596. [PREPRINT]
- Csordas, R., Potts, C., Manning, C. D., & Geiger, A. (2024). Recurrent Neural Networks Learn to Store and Generate Sequences using Non-Linear Representations. arXiv:2408.10920. doi:10.48550/arXiv.2408.10920. [PREPRINT]
- Cui, J., Zhang, Q., Wang, Y., & Wang, Y. (2025). On the Limits of Sparse Autoencoders: A Theoretical Framework and Reweighted Remedy. ICLR 2026. arXiv:2506.15963. doi:10.48550/arXiv.2506.15963.
- Cunningham, H., Ewart, A., Riggs, L., Huben, R., & Sharkey, L. (2023). Sparse Autoencoders Find Highly Interpretable Features in Language Models. arXiv:2309.08600. doi:10.48550/arXiv.2309.08600.
- Dunefsky, J., Chlenski, P., & Nanda, N. (2024). Transcoders Find Interpretable LLM Feature Circuits. NeurIPS 2024. arXiv:2406.11944. doi:10.48550/arXiv.2406.11944.
- Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., et al. (2022). Toy Models of Superposition. arXiv:2209.10652. doi:10.48550/arXiv.2209.10652.
- Engels, J., Michaud, E. J., Liao, I., Gurnee, W., & Tegmark, M. (2024a). Not All Language Model Features Are One-Dimensionally Linear. ICLR 2025. arXiv:2405.14860. doi:10.48550/arXiv.2405.14860.
- Engels, J., Riggs, L., & Tegmark, M. (2024b). Decomposing The Dark Matter of Sparse Autoencoders. Transactions on Machine Learning Research. arXiv:2410.14670. doi:10.48550/arXiv.2410.14670.
- Enkhbayar, T. (2025). Which Sparse Autoencoder Features Are Real? Model-X Knockoffs for False Discovery Rate Control. arXiv:2511.11711. doi:10.48550/arXiv.2511.11711. [PREPRINT]
- Erdogan, E., & Lucic, A. (2025). Equivariant Sparse Autoencoders: Mechanistic Interpretability of Neural Networks on Symmetric Data. arXiv:2511.09432. doi:10.48550/arXiv.2511.09432. [PREPRINT]
- Farrell, E., Lau, Y.-T., & Conmy, A. (2024). Applying sparse autoencoders to unlearn knowledge in language models. arXiv:2410.19278. doi:10.48550/arXiv.2410.19278. [PREPRINT]
- Fel, T., Lubana, E. S., Prince, J. S., Kowal, M., Boutin, V., Papadimitriou, I., Wang, B., Wattenberg, M., Ba, D., & Konkle, T. (2025). Archetypal SAE: Adaptive and Stable Dictionary Learning for Concept Extraction in Large Vision Models. ICML 2025. arXiv:2502.12892. doi:10.48550/arXiv.2502.12892.
- Gao, L., Dupre la Tour, T., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., & Wu, J. (2024). Scaling and evaluating sparse autoencoders. arXiv:2406.04093. doi:10.48550/arXiv.2406.04093.
- Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., Arora, A., Wu, Z., Goodman, N., Potts, C., & Icard, T. (2023). Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability. arXiv:2301.04709. doi:10.48550/arXiv.2301.04709.
- Gerasimov, G., Rusalev, T., Balagansky, N., Laptev, D., Kurochkin, V., & Gavrilov, D. (2026). Unstable Features, Reproducible Subspaces: Understanding Seed Dependence in Sparse Autoencoders. arXiv:2606.12138. doi:10.48550/arXiv.2606.12138. [PREPRINT]
- He, Z., Shu, W., Ge, X., Chen, L., Wang, J., Zhou, Y., Liu, F., Guo, Q., Huang, X., Wu, Z., Jiang, Y.-G., & Qiu, X. (2024). Llama Scope:

- Extracting Millions of Features from Llama-3.1-8B with Sparse Autoencoders. arXiv:2410.20526. doi:10.48550/arXiv.2410.20526. [PREPRINT]
- Heap, T., Lawson, T., Farnik, L., & Aitchison, L. (2025). Automated Interpretability Metrics Do Not Distinguish Trained and Random Transformers. arXiv:2501.17727. doi:10.48550/arXiv.2501.17727. [PREPRINT]
- Hillar, C. J., & Sommer, F. T. (2011). When can dictionary learning uniquely recover sparse data from subsamples? arXiv:1106.3616. doi:10.48550/arXiv.1106.3616.
- Hindupur, S. S. R., Lubana, E. S., Fel, T., & Ba, D. (2025). Projecting Assumptions: The Duality Between Sparse Autoencoders and Concept Geometry. NeurIPS 2025. arXiv:2503.01822. doi:10.48550/arXiv.2503.01822.
- Hoang, P. G., Chatterjee, A., Chakraborty, T., Gurevych, I., & Dutta, S. (2026). Sparse Autoencoders Encode Both Concepts and Functions: The Downstream Geometry of Feature Effects. arXiv:2607.24645. doi:10.48550/arXiv.2607.24645. [PREPRINT]
- Huang, J., Wu, Z., Potts, C., Geva, M., & Geiger, A. (2024). RAVEL: Evaluating Interpretability Methods on Disentangling Language Model Representations. ACL 2024. arXiv:2402.17700. doi:10.48550/arXiv.2402.17700.
- Joshi, S., Dittadi, A., Lachapelle, S., & Sridhar, D. (2025). Sparse Shift Autoencoders for Identifying Concepts from Large Language Model Activations. arXiv:2502.12179. doi:10.48550/arXiv.2502.12179. [PREPRINT]
- Kantamneni, S., Engels, J., Rajamanoharan, S., Tegmark, M., & Nanda, N. (2025). Are Sparse Autoencoders Useful? A Case Study in Sparse Probing. arXiv:2502.16681. doi:10.48550/arXiv.2502.16681. [PREPRINT]
- Karvonen, A., Rager, C., Lin, J., Tigges, C., Bloom, J., Chanin, D., Lau, Y.-T., Farrell, E., McDougall, C., Ayonrinde, K., Till, D., Wearden, M., et al. (2025). SAE-Bench: A Comprehensive Benchmark for Sparse Autoencoders in Language Model Interpretability. ICML 2025. arXiv:2503.09532. doi:10.48550/arXiv.2503.09532.
- Karvonen, A., Wright, B., Rager, C., Angell, R., Brinkmann, J., Smith, L., Mayrink Verdun, C., Bau, D., & Marks, S. (2024). Measuring Progress in Dictionary Learning for Language Model Interpretability with Board Game Models. NeurIPS 2024. arXiv:2408.00113. doi:10.48550/arXiv.2408.00113.
- Khemakhem, I., Kingma, D. P., Monti, R. P., & Hyvarinen, A. (2020). Variational Autoencoders and Nonlinear ICA: A Unifying Framework. AISTATS 2020. arXiv:1907.04809. doi:10.48550/arXiv.1907.04809.
- Korznikov, A., Galichin, A., Dontsov, A., Rogov, O., Tutubalina, E., & Oseledets, I. (2025). OrtsAE: Orthogonal Sparse Autoencoders Uncover Atomic Features. arXiv:2509.22033. doi:10.48550/arXiv.2509.22033. [PREPRINT]
- Lachapelle, S., Rodriguez Lopez, P., Sharma, Y., Everett, K., Le Priol, R., Lacoste, A., & Lacoste-Julien, S. (2022). Disentanglement via Mechanism Sparsity Regularization: A New Principle for Nonlinear ICA. CLeaR 2022. arXiv:2107.10098. doi:10.48550/arXiv.2107.10098.
- Lan, M., Torr, P., Meek, A., Khakzar, A., Krueger, D., & Barex, F. (2024). Quantifying Feature Space Universality Across Large Language Models via Sparse Autoencoders. arXiv:2410.06981. doi:10.48550/arXiv.2410.06981. [PREPRINT]
- Leask, P., Busmann, B., Pearce, M., Bloom, J., Tigges, C., Al Moubayed, N., Sharkey, L., & Nanda, N. (2025). Sparse Autoencoders Do Not Find Canonical Units of Analysis. ICLR 2025. arXiv:2502.04878. doi:10.48550/arXiv.2502.04878.
- Li, Y., Michaud, E. J., Baek, D. D., Engels, J., Sun, X., & Tegmark, M. (2024). The Geometry of Concepts: Sparse Autoencoder Feature Structure. Entropy, 27(4), 344 (2025). arXiv:2410.19750. doi:10.48550/arXiv.2410.19750.
- Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramar, J., Dragan, A., Shah, R., & Nanda, N. (2024). Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2. arXiv:2408.05147. doi:10.48550/arXiv.2408.05147.
- Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schoelkopf, B., & Bachem, O. (2019). Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. ICML 2019. arXiv:1811.12359. doi:10.48550/arXiv.1811.12359.
- Makelov, A., Lange, G., & Nanda, N. (2023). Is This the Subspace You Are Looking for? An Interpretability Illusion for Subspace Activation Patching. arXiv:2311.17030. doi:10.48550/arXiv.2311.17030. [PREPRINT]
- Makelov, A., Lange, G., & Nanda, N. (2024). Towards Principled Evaluations of Sparse Autoencoders for Interpretability and Control. arXiv:2405.08366. doi:10.48550/arXiv.2405.08366. [PREPRINT]
- Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., & Mueller, A. (2024). Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models. ICLR 2025. arXiv:2403.19647. doi:10.48550/arXiv.2403.19647.
- Marks, S., Treutlein, J., Bricken, T., Lindsey, J., Marcus, J., Mishra-Sharma, S., Ziegler, D., Ameisen, E., Batson, J., Belonax, T., Bowman, S. R., Carter, S., et al. (2025). Auditing language models for hidden objectives. arXiv:2503.10965. doi:10.48550/arXiv.2503.10965. [PREPRINT]
- Menon, A., Shrivastava, M., Krueger, D., & Lubana, E. S. (2024). Analyzing (In)Abilities of SAEs via Formal Languages. NAACL 2025. arXiv:2410.11767. doi:10.48550/arXiv.2410.11767.
- Michaud, E. J., Gorton, L., & McGrath, T. (2025). Understanding sparse autoencoder scaling in the presence of feature manifolds. arXiv:2509.02565. doi:10.48550/arXiv.2509.02565. [PREPRINT]
- Minder, J., Dumas, C., Juang, C., Chughtai, B., & Nanda, N. (2025). Overcoming Sparsity Artifacts in Crosscoders to Interpret Chat-Tuning. NeurIPS 2025. arXiv:2504.02922. doi:10.48550/arXiv.2504.02922.
- Movva, R., Peng, K., Garg, N., Kleinberg, J., & Pierson, E. (2025). Sparse Autoencoders for Hypothesis Generation. ICML 2025. arXiv:2502.04382. doi:10.48550/arXiv.2502.04382.
- Mueller, A., Geiger, A., Wiegrefe, S., Arad, D., Arcuschin, I., Belfki, A., Chan, Y. S., Fiotto-Kaufman, J., Haklay, T., Hanna, M., Huang, J., Gupta, R., et al. (2025). MIB: A Mechanistic Interpretability Benchmark. ICML 2025. arXiv:2504.13151. doi:10.48550/arXiv.2504.13151.
- Muhamed, A., Bonato, J., Diab, M., & Smith, V. (2025). SAEs Can Improve Unlearning: Dynamic Sparse Autoencoder Guardrails for Precision Unlearning in LLMs. COLM 2025. arXiv:2504.08192. doi:10.48550/arXiv.2504.08192.
- Nelson, W., Karaletsos, T., & Locatello, F. (2026). Toward Identifiable Sparse Autoencoders. ICML 2026. arXiv:2605.31245. doi:10.48550/arXiv.2605.31245.
- Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 381, 607-609. doi:10.1038/381607a0.
- O'Neill, C., Gumran, A., & Klindt, D. (2025). Compute Optimal Inference and Provable Amortisation Gap in Sparse Autoencoders. arXiv:2411.13117. doi:10.48550/arXiv.2411.13117. [PREPRINT]
- Park, K., Choe, Y. J., & Veitch, V. (2024). The Linear Representation Hypothesis and the Geometry of Large Language Models. ICML 2024. arXiv:2311.03658. doi:10.48550/arXiv.2311.03658.
- Paulo, G., & Belrose, N. (2025). Sparse Autoencoders Trained on the Same Data Learn Different Features. arXiv:2501.16615. doi:10.48550/arXiv.2501.16615. [PREPRINT]
- Paulo, G., Mallen, A., Juang, C., & Belrose, N. (2024). Automatically Interpreting Millions of Features in Large Language Models. arXiv:2410.13928. doi:10.48550/arXiv.2410.13928. [PREPRINT]
- Paulo, G., Shabalin, S., & Belrose, N. (2025). Transcoders Beat Sparse Autoencoders for Interpretability. arXiv:2501.18823. doi:10.48550/arXiv.2501.18823. [PREPRINT]
- Peng, K., Movva, R., Kleinberg, J., Pierson, E., & Garg, N. (2025). Position: Use Sparse Autoencoders to Discover Unknowns. ICML

2026. arXiv:2506.23845. doi:10.48550/arXiv.2506.23845.
- Rajamanoharan, S., Conmy, A., Smith, L., Lieberum, T., Varma, V., Kramar, J., Shah, R., & Nanda, N. (2024a). Improving Dictionary Learning with Gated Sparse Autoencoders. arXiv:2404.16014. doi:10.48550/arXiv.2404.16014.
- Rajamanoharan, S., Lieberum, T., Sonnerat, N., Conmy, A., Varma, V., Kramar, J., & Nanda, N. (2024b). Jumping Ahead: Improving Reconstruction Fidelity with JumpReLU Sparse Autoencoders. arXiv:2407.14435. doi:10.48550/arXiv.2407.14435.
- Saphra, N., & Wiegrefe, S. (2024). Mechanistic? arXiv:2410.09087. doi:10.48550/arXiv.2410.09087.
- Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., Biderman, S., Garriga-Alonso, A., et al. (2025). Open Problems in Mechanistic Interpretability. arXiv:2501.16496. doi:10.48550/arXiv.2501.16496. [PREPRINT]
- Simon, E., & Zou, J. (2024). InterPLM: Discovering Interpretable Features in Protein Language Models via Sparse Autoencoders. arXiv:2412.12101. doi:10.48550/arXiv.2412.12101. [PREPRINT]
- Song, X., Muhamed, A., Zheng, Y., Kong, L., Tang, Z., Diab, M. T., Smith, V., & Zhang, K. (2025). Position: Mechanistic Interpretability Should Prioritize Feature Consistency in SAEs. arXiv:2505.20254. doi:10.48550/arXiv.2505.20254. [PREPRINT]
- Spielman, D. A., Wang, H., & Wright, J. (2012). Exact Recovery of Sparsely-Used Dictionaries. arXiv:1206.5882. doi:10.48550/arXiv.1206.5882.
- Sun, X., Stolfo, A., Engels, J., Wu, B., Rajamanoharan, S., Sachan, M., & Tegmark, M. (2025). Dense SAE Latents Are Features, Not Bugs. NeurIPS 2025. arXiv:2506.15679. doi:10.48550/arXiv.2506.15679.
- Thasarathan, H., Forsyth, J., Fel, T., Kowal, M., & Derpanis, K. G. (2025). Universal Sparse Autoencoders: Interpretable Cross-Model Concept Alignment. arXiv:2502.03714. doi:10.48550/arXiv.2502.03714. [PREPRINT]
- Venhoff, C., Calinescu, A., Torr, P., & Schroeder de Witt, C. (2024). SAGE: Scalable Ground Truth Evaluations for Large Sparse Autoencoders. arXiv:2410.07456. doi:10.48550/arXiv.2410.07456. [PREPRINT]
- Vilas, M. G., Adolphi, F., Poeppel, D., & Roig, G. (2024). Position: An Inner Interpretability Framework for AI Inspired by Lessons from Cognitive Neuroscience. ICML 2024. arXiv:2406.01352. doi:10.48550/arXiv.2406.01352.
- Williams, I., Oldenburg, N., Dhar, R., Hatherley, J., Fierro, C., Rajcic, N., Schiller, S. R., Stamatiou, F., & Sogaard, A. (2025). Mechanistic Interpretability Needs Philosophy. arXiv:2506.18852. doi:10.48550/arXiv.2506.18852. [PREPRINT]
- Wu, Z., Arora, A., Geiger, A., Wang, Z., Huang, J., Jurafsky, D., Manning, C. D., & Potts, C. (2025). AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders. arXiv:2501.17148. doi:10.48550/arXiv.2501.17148. [PREPRINT]