WHEN DO SAFETY-RELEVANT CONCEPTS BECOME READABLE IN THE TRAINING PROCESS OF AN LLM?

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models in everyday products are expected to avoid toxic, harmful, or stereotyped text. The concepts behind such text take shape during pretraining, so monitoring training requires knowing when each appears. Existing studies follow a concept across saved checkpoints with the raw score of a *readout*, a method that separates texts with and without the concept using the model's activations: a linear probe or a few sparse autoencoder (SAE) features. However, a raw score cannot tell when the concept became readable, because the two kinds of text differ in their words and the score can be high before any learning. We show that this point in training can be found, and that raw scores miss it. We measure every readout against its own score at the first checkpoint, when the model is still nearly random, and date a concept when this gain stays above a threshold. On 32 concepts across the checkpoints of our primary model, the raw score of a linear probe would date nine to the start of training. We instead find that toxicity emerges first, early in pretraining (by step 1,000). At our threshold, harm categories, which mark the kind of harm a text describes, emerge later and only under the SAE readout, and stereotype (bias) concepts under neither. The date thus depends on the readout, which a single-readout study cannot see. At the final checkpoint, removing a few SAE feature directions reduces generated toxicity. A set of five directions fixed in advance passes checks on benign prompts and output diversity, replicating in every pre-registered combination of three behaviours (toxicity, insult, threat) and two sizes of the primary model (1.4B and 2.8B).

1 INTRODUCTION

Language models now write and converse for a broad public and are expected to do so without toxic, harmful, or stereotyped responses (Gehman et al., 2020; Ji et al., 2023; Nangia et al., 2020). The representations behind such behaviour are built during pretraining, yet they are usually examined only after pretraining has ended. Public checkpoint series such as Pythia (Biderman et al., 2023)[1] and OLMo-2 (OLMo Team et al., 2025)[2] make it possible to look inside a model while it is still being trained and to ask at what point a safety-relevant concept becomes visible in its activations. A rising score is informative only if it reflects learning rather than the evaluation texts or the reading method. This paper asks *when safety-relevant concepts become readable during pretraining, and how the answer depends on how one reads*. We call a method that reads a concept from activations a *readout*, and we call a concept *readable* at a checkpoint when a readout distinguishes texts that carry the concept from texts that do not. Readable does not mean that generated text is easy to read or that the model is safe. A linear probe is trained with labels to separate the two kinds of text, whereas a sparse autoencoder (SAE) learns a dictionary of features without labels, from which a few are selected to separate them. Either readout yields directions in activation space; removing or adding such directions, called steering, tests whether the model relies on them.

Prior checkpoint studies track such scores over training (Ge et al., 2026; Xu et al., 2024; Inaba et al., 2025), but four difficulties, D1 to D4, stand between a score and a claim about learning. The step at which a concept appears may depend on the readout (**D1**). Many safety benchmarks consist of minimal pairs, two sentences that differ in a few tokens and carry opposite labels, so a probe trained on one sentence can assign the same label to the other sentence of the pair (its mate) at evaluation, and its score can fall below chance (**D2**). A high early score may come from surface token statistics that near-random weights preserve (**D3**). Finally, reading a concept does not tell us whether removing it changes what the model generates (**D4**).

[1] https://github.com/EleutherAI/pythia

[2] https://huggingface.co/allenai/OLMo-2-0425-1B

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We address these difficulties with a controlled audit of probes and SAEs across pretraining (Figure 1). For D1, four readouts (a linear probe, a mean-difference direction, and SAE top-1 and top-5) run on identical folds, so their dates are comparable. For D2, every minimal pair stays in one fold, so no probe is evaluated on the mate of a training text. For D3, each readout's trajectory is measured against its own first-checkpoint score, its *lexical floor*: one gradient step from random weights, little beyond surface token statistics separates the texts. A concept whose floor is already high is flagged as separable before learning; one whose corrected score rises above a threshold and holds at the next checkpoint is dated to that checkpoint. For D4, a separate experiment at the final checkpoint removes the directions a readout selects and tests, against random directions of the same rank, whether the concept stays readable and whether generated text changes.

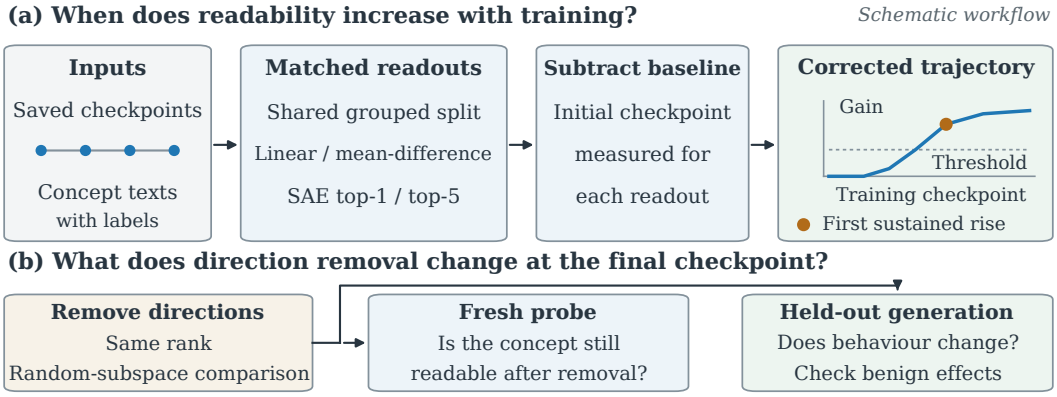

Figure 1: **From checkpoints to a concept timeline and an intervention test.** (a) Shared folds and a first-checkpoint baseline turn checkpoints into a corrected trajectory; the orange point marks the first sustained rise. (b) Removing the selected directions at the final checkpoint shows whether the model relies on them: a refit probe tests remaining readability, held-out generation tests behaviour.

Five experiments organise the evaluation (§5): E1 compares readouts (D1, D2); E2 measures lexical floors (D3); E3 dates sustained increases (D1, D3); E4 checks robustness to seeds, architectures, and scale; E5 tests what removing directions changes for probes and generated text (D4).
Our main claim is that when a concept becomes readable is well defined only for a stated readout, whichever of the four it is: the date must come with that readout's own first-checkpoint score (its lexical floor) and, for an SAE readout, with the dictionary's training budget (§3 and §4). On 32 concepts in Pythia-1.4B (§5; §5.4 tests larger Pythia models and OLMo-2) we find:

- Linear probes read best among the four readouts (E1).
- Under the linear probe, nine concepts are already separable at the first checkpoint (E2).
- After correction, the toxicity group becomes readable first (by step 1,000). At a 0.15 gain threshold, the harm-category group follows only under the SAE readout and the bias group under neither, and the dictionaries' training budget moves most of the SAE concept-group dates (E3).
- Among steering edits of at most five directions, SAE directions give the deepest reductions within the pre-registered criteria in all but one of the six settings (three behaviours, Pythia-1.4B and 2.8B), yet directions selected for unrelated concepts cut toxicity as much at 1.4B (E5).

We can thus tell whether a concept is readable beyond surface statistics, whether that depends on the readout, and what removing directions changes in generated text. A raw score under one readout tells none of these. For the first two, Proposition 1 shows this in theory and E1–E3 in the data.

2 Related Work

Sparse autoencoders decompose residual activations into interpretable feature dictionaries (Bricken et al., 2023; Huben et al., 2024), including TopK (Gao et al., 2025). They have been scaled to frontier models (Templeton et al., 2024), benchmarked (Karvonen et al., 2025; Huang et al., 2024), and learn seed-dependent dictionaries (Paulo & Belrose, 2026). Concept-aligned dictionaries supervise features toward given concepts (Yang et al., 2025); we use labels only to select features. Whether SAE features beat raw activations as probes has been studied on trained models with mostly negative or mixed answers (Kantamneni et al., 2025; Gallifant et al., 2025); we find the same and extend the comparison along the trajectory and to what removing selected features changes.

Table 1: **Which questions can each line of work answer?** Traj.: tracks concepts across pretraining checkpoints; Readouts: compares several readouts on one labelled contrast; Floor: checks separability against an untrained or control baseline; On one trajectory: the Readouts and Floor checks made along one checkpoint trajectory; Edit: measures what removing directions changes; Seeds, Arch., Scale: repeat a central result across seeds, model families, and sizes; a check marks a central focus of at least one cited work.

	Dating concepts			On one trajectory		Robustness			
	Traj.	Readouts	Floor	Readouts	Floor	Edit	Seeds	Arch.	Scale
Control-task probing (Hewitt & Liang, 2019; Zhang & Bowman, 2018)	–	✓	✓	–	–	–	–	–	–
SAE dictionaries (Gao et al., 2025; Templeton et al., 2024; Paulo & Belrose, 2026)	–	✓	–	–	–	✓	✓	✓	✓
SAE probing (Kantamneni et al., 2025; Gallifant et al., 2025)	–	✓	–	–	–	–	–	✓	✓
SAE trajectory studies (Ge et al., 2026; Xu et al., 2024)	✓	–	–	–	–	✓	✓	✓	✓
Direction ablation (Arditi et al., 2024; Belrose et al., 2023)	–	–	–	–	–	✓	–	✓	✓
Ours	✓	✓	✓	✓	✓	✓	✓	✓	✓

Classic probing showed that probe accuracy conflates what a representation encodes with what a probe can extract from any input, and introduced control tasks (Hewitt & Liang, 2019) and random-encoder baselines (Zhang & Bowman, 2018) to separate the two; we apply this distinction along a trajectory with a first-checkpoint floor. Stereotype benchmarks are minimal pairs whose two sides differ in a few tokens (Nangia et al., 2020), with contested validity as bias measures (Blodgett et al., 2021); we keep each pair on one side of every split.

Pythia (Biderman et al., 2023) and OLMo-2 (OLMo Team et al., 2025) release dense intermediate checkpoints. The studies closest to ours track feature formation across such checkpoints with a single readout family each (Ge et al., 2026; Xu et al., 2024; Inaba et al., 2025), and Ge et al. (2026) also connect features to downstream tasks by ablation; none dates a labelled concept against its first-checkpoint score under more than one readout, so they cannot tell whether a date belongs to the concept or to the readout. Universal-neuron studies (Gurnee et al., 2024) and metric-dependent emergence (Wei et al., 2022; Schaeffer et al., 2023; Nanda et al., 2023) show that a measured onset can reflect the measure or surface statistics rather than learning, which we address with floors and a two-checkpoint requirement.

On the safety side, refusal has been isolated as a single removable direction in aligned chat models (Arditi et al., 2024), concept-erasure methods remove linear structure (Belrose et al., 2023; Ravfogel et al., 2020), and activation addition translates along a mean-difference direction (Turner et al., 2023; Rimsky et al., 2024). Attention-head interventions (Li et al., 2023), representation engineering (Zou et al., 2023a), and SAE feature control (Makelov et al., 2024; Marks et al., 2025; Wang et al., 2026) extend the recipe. Decoding-time detoxification (Liu et al., 2021; Krause et al., 2021) reweights next-token probabilities and edits no activation direction, so it cannot test whether the model relies on a readout's directions (D4); we therefore compare representation-level methods only (§5.5).

Table 1 states what each column lets one decide. The trajectory makes change visible. The floor shows whether labels were separable before learning. A second readout on the same folds shows whether a date belongs to the concept or to the reading method, and only the three together on one trajectory attribute a date to learning. The edit test shows whether the model relies on the selected directions, and the robustness columns whether a result survives other seeds, architectures, and sizes.

3 A Controlled Audit of Readability Across Checkpoints

3.1 Problem Setup and Notation

We are given the saved checkpoints of one pretraining run and, for each concept, texts labelled as carrying the concept or not. We want to know at which checkpoint a concept becomes readable beyond what was already separable before learning, whether that checkpoint changes with the readout, and whether the model relies on the directions that a readout selects. For every concept, readout, and checkpoint we therefore compute a held-out classification score, take its value at the first checkpoint as a baseline, and date the concept to the first checkpoint at which the gain over this baseline stays above a threshold. At the final checkpoint we remove the selected directions and measure what changes. Algorithm 1 summarises these steps with the symbols that the rest of this section defines.

The inputs of Algorithm 1 are the checkpoints, the labelled texts of each concept, and the readouts. Texts that share a provenance form one *source group*, such as the two sentences of a minimal pair or the texts built from one question, and an unrelated text forms a group of its own. A *readout* fits a scoring function on labelled activations and scores unseen texts, and we measure its discrimination by the area under the receiver operating characteristic curve (AUROC).

Algorithm 1 Measure a concept timeline, then test an intervention.

Require: Checkpoints M_t ($t \in \mathcal{T}$), layer ℓ, labelled texts $\mathcal{D}_c$ with source groups g_c for each concept c, readouts R, threshold θ, rank m

1: For every c, R, and t: $A_c^R(t) \leftarrow$ held-out AUROC of R on $\mathcal{D}_c$ at layer ℓ of M_t, with each source group kept in one fold (§3.1, §3.2).
2: Lexical floor: $F_c^R \leftarrow A_c^R(t_0)$, the score at the first checkpoint t_0 (§3.3).
3: Floor-corrected trajectory: $\Delta_c^R(t) \leftarrow A_c^R(t) - F_c^R$.
4: Emergence step: $\tau_c^R(\theta) \leftarrow$ the first t with $\Delta_c^R > \theta$ at t and at the next checkpoint.
5: At the final checkpoint, remove the rank-m subspace Q of the selected directions; measure the AUROC lost by a fixed probe, $\sigma_{\text{fix}}(Q)$, and by a refit probe, $\sigma_{\text{fresh}}(Q)$, and the change in generated text (§3.4).
6: **return** $A_c^R(t)$, F_c^R, and $\tau_c^R(\theta)$ for every c and R; $\sigma_{\text{fix}}(Q)$, $\sigma_{\text{fresh}}(Q)$, and the generation effects.

In Algorithm 1, M_t is the model checkpoint after t pretraining steps, on a finite grid $t \in \mathcal{T}$ with earliest point t_0. The linear probe, the mean-difference direction, and Theorem 1 are defined on $\phi_\ell(x; M_t) \in \mathbb{R}^d$, the residual-stream activation at layer ℓ (the hidden state after that layer), mean-pooled over the tokens of text x (ℓ and M_t are suppressed when fixed). Each concept c has balanced labelled texts $\mathcal{D}_c = \{(x_i, y_i)\}_{i=1}^{n_c}$, its *probe set*, with binary labels (§4), and $g_c(i) \in \mathcal{G}_c$ records the source group of text i, frozen before extraction. A readout R gives text x_i the score s_i. For a held-out evaluation set $\mathcal{E}$ (texts unseen in fitting) with $n^+ \geq 1$ positive and $n^- \geq 1$ negative examples, the AUROC $\mathrm{A}(s; \mathcal{E})$ of the scores s is

$$\mathrm{A}(s; \mathcal{E}) = \frac{1}{n^+ n^-} \sum_{i: y_i = 1} \sum_{j: y_j = 0} \Big(\mathbf{1}\big[s_i > s_j\big] + \tfrac{1}{2}\, \mathbf{1}\big[s_i = s_j\big] \Big) \;\in\; [0, 1]. \tag{1}$$

This is the fraction of positive–negative pairs in which the concept-positive text receives the higher score, counting a tie as half. An AUROC of 0.5 is chance; higher values mean better separation.

3.2 Comparing Readouts Without Splitting Related Texts

Step 1 of Algorithm 1 compares four readouts on identical data: a linear probe, a mean-difference direction, and two SAE readouts, top-1 and top-5. The linear probe is ℓ_2-regularised logistic regression on activations. The mean-difference direction projects activations onto the difference of the positive and negative class means. For the SAE readouts, a separate TopK SAE (keeping the k=32 largest feature activations per token) is trained at each checkpoint on unlabelled text from the Pile (Gao et al., 2020); it encodes each token, and the feature activations are averaged over the text. Each feature also has a decoder direction, the column of the SAE decoder that reconstructs activations from that feature, so selected features define directions in activation space. The SAE top-1 and top-5 readouts (the readout size, distinct from k) select the one or five features whose training-fold AUROC is farthest from chance, orient each toward the positive class, and average their signed activations.

All readouts keep source groups intact and confine supervised fitting and feature selection to training folds, so evaluation never sees the mate of a training text and evaluation labels never enter feature choice. A row-level split instead assigns each text to a fold on its own and can separate the two sentences of a minimal pair, and Theorem 1 shows that a probe can then score below chance.

Definition 1 (Grouped audited readout)**.** A partition of the labelled set $\mathcal{D}_c$ into $K \geq 2$ disjoint folds $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(K)}$ is *group-respecting* if texts of one source group, $g_c(i) = g_c(j)$, share a fold. The *audited AUROC* $A_c^R(t)$ of readout R at checkpoint t is the mean over the K folds of the held-out AUROC of a scorer $s^{(-q)}$ produced from $\mathcal{D}_c \setminus \mathcal{D}^{(q)}$ alone (Eq. 7, Appendix A): every supervised fitting and feature-selection operation, including SAE ranking and orientation, uses training folds only.

Folds are stratified over groups with K=4 and fixed across readouts, checkpoints, and interventions, so that scores are comparable. Theorem 1 states the failure precisely: a text closer to its training mate than the probe's margin is scored on the wrong side.

Theorem 1 (Minimal-pair inversion under row-level folds)**.** *Let $\mathcal{D}_c$ consist of minimal pairs with opposite labels and let a row-level fold assignment leave sets $M^\pm$ of evaluation items whose mates lie in the training fold, with mated fractions $\pi^\pm = |M^\pm|/n^\pm$. Suppose mates are δ-close, $\|\phi(x_\gamma^+) - \phi(x_\gamma^-)\| \leq \delta$ for every pair γ, and the trained scorer $s(\phi) = \langle w, \phi \rangle + b$ with $w \neq 0$ separates its training fold with geometric margin $\eta > \delta$, that is $y_i\big(\langle w, \phi_i \rangle + b\big) \geq \eta \|w\|$ with labels coded ± 1. Then every mated evaluation pair of opposite classes is ordered backwards, and $\mathrm{A} \leq 1 - \pi^+ \pi^-$.*

Under this margin condition the bound falls below chance whenever $\pi^+ \pi^- > 1/2$; grouped folds set $\pi^\pm = 0$, so the bound is vacuous (Corollary 1 and the full proof, Appendix B.1).

3.3 Dating Increases Above the Initial Baseline

A high score at an early checkpoint need not mark a newly learned representation, because the texts may already be separable through token statistics that near-random weights preserve. We therefore measure each readout at the earliest checkpoint and track its later increase over that baseline, the lexical floor, an operational baseline rather than an exact lexical/learned split.

Definition 2 (Lexical floor and emergence step)**.** The *lexical floor* F_c^R of c under R is

$$F_c^R \;=\; A_c^R(t_0), \qquad t_0 = \min \mathcal{T}, \tag{2}$$

the audited AUROC at the earliest checkpoint (step 1 on Pythia, step 0 on OLMo-2), at most one gradient step from random weights. The *floor-corrected* trajectory is $\Delta_c^R(t) = A_c^R(t) - F_c^R$, and c is *flagged* under R if $F_c^R > 0.8$. For a threshold $\theta > 0$, with the next grid point denoted by $\mathrm{succ}(t)$ and $\mathcal{T}^- = \mathcal{T} \setminus \{\max \mathcal{T}\}$, the corrected *emergence step* $\tau_c^R(\theta)$ is

$$\tau_c^R(\theta) = \inf\big\{t \in \mathcal{T}^- : \min\big(\Delta_c^R(t), \Delta_c^R(\mathrm{succ}(t))\big) > \theta\big\}, \tag{3}$$

and the raw step $\tilde{\tau}_c^R(a)$ at a level $a \in [0,1]$ is the first checkpoint with $A_c^R(t) \geq a$. We set $\inf \emptyset = \infty$; two successive scores above θ discard isolated spikes; reported timelines use $\theta = 0.15$, fixed before the constant-budget experiment and varied in §5.3. A concept group (related concepts, §4) applies the same rule to the median of $\Delta_c^R(t)$ over its concepts, which gives the group step.

Proposition 1 records three consequences for reading timelines (proof in Appendix B.2). The corrected gain cannot exceed the headroom $1 - F_c^R$ (i). A raw threshold no higher than the floor is met at the first checkpoint without later learning (ii), so a raw step is uninformative for a flagged concept. Once θ reaches the headroom no crossing is possible (iii), so an absent crossing does not show the concept absent.

Proposition 1 (Limits that the floor places on raw and corrected steps)**.** *For all c, R, t, $a \in [0,1]$, $\theta > 0$:* (i) $\Delta_c^R(t) \leq 1 - F_c^R$; (ii) *if* $a \leq F_c^R$ *then* $\tilde{\tau}_c^R(a) = t_0$; (iii) $\theta \geq 1 - F_c^R$ *forces* $\tau_c^R(\theta) = \infty$.

3.4 Testing What Removing Directions Changes

Readability measures label information that a readout can access, not whether the model uses the selected directions when generating. We remove them at the final checkpoint and ask: does the original probe lose AUROC, can a refit probe recover it, and do independently scored generations change.

Definition 3 (Subspace removal and its sensitivities)**.** At the final checkpoint let $Q \in \mathbb{R}^{d \times m}$ be an orthonormal basis of the rank-m subspace to be removed, spanned by the selected SAE decoder directions or by the supervised or random comparison directions. The removal rank is chosen separately from the one or five features used for reading. Removal projects every activation onto the orthogonal complement of Q (Eq. 8, Appendix A). The *fixed-probe sensitivity* $\sigma_{\mathrm{fix}}(Q)$ is the held-out AUROC that a probe fit on unmodified training-fold activations loses when Q is removed at evaluation; the *fresh-probe sensitivity* $\sigma_{\mathrm{fresh}}(Q)$ is the loss against the same unedited AUROC after the probe is refit on training-fold activations with Q removed (Eq. 9). Both are averaged over folds, with Q and the probes computed per fold as in Definition 1, and positive values are AUROC lost. The behavioural effect is the change in independently scored generation statistics when the projection is applied to each token's residual during prompt processing and decoding.

If the fixed probe loses AUROC but a refit probe regains it (large σ_{fix}, small σ_{fresh}), the concept is still linearly readable after the edit, and the loss reflects the overlap of the fixed probe's weights with Q. A loss that survives refitting (large σ_{fresh}) suggests that readable information was removed. Proposition 2 (Appendix A, proof in Appendix B.3) formalises the first reading at population level. Random subspaces of the same rank calibrate the projection: one removes an expected m/d of a fixed probe direction's squared norm (Proposition 3, Appendix A, proof in Appendix B.4), and ten random subspaces give the reference for generation, which has no closed form.

4 Experimental Setup

Our primary trajectory is Pythia-1.4B, 20 log-spaced public checkpoints from step 1 to 143,000, analysed at layers 12 and 18 (of 24): the middle of the stack as the primary layer and a later layer for robustness. The same recipe runs on 11 log-spaced OLMo-2-1B stage-1 checkpoints (0–4,001B tokens) at layers 6 and 12 (of 16), chosen to roughly match those depths, and, for scale, on Pythia-2.8B, 6.9B, and 12B at three checkpoints at a mid-depth layer. All SAEs are TopK autoencoders at $8\times$ expansion trained on Pile text, with three dictionary seeds per Pythia-1.4B checkpoint and layer (42, then 123 and 7). Every layer-12 dictionary at steps up to 1,000, trained on 10^6 to 3.33×10^6 tokens under the original schedule, was also retrained for all three seeds on the late checkpoints' budget of 10^7 tokens; these constant-budget dictionaries form the reported layer-12 trajectory (Appendix F.1).

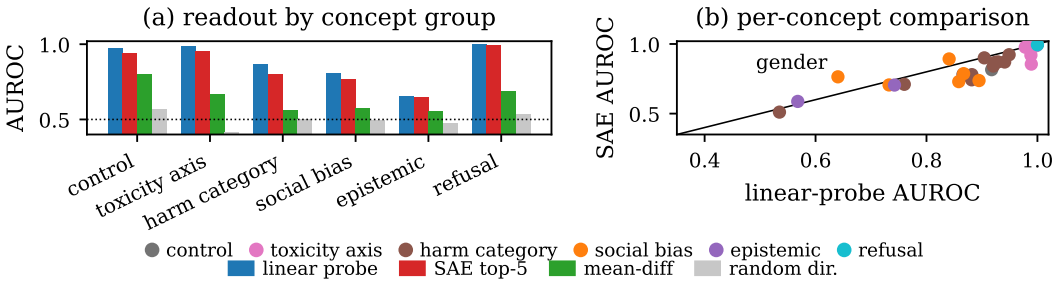

Figure 2: **Readouts at the final checkpoint.** (a) Mean AUROC per group. (b) SAE top-5 (seed 42) against linear-probe AUROC per concept; above the diagonal favours the SAE.

We study 32 binary concepts in six concept groups; a concept group collects related concepts, unlike a source group (§3.1), which keeps related texts in one fold. Thirteen are BeaverTails harm categories (Ji et al., 2023), each contrasted with texts from *other* harm categories, so the label isolates category identity rather than harmfulness (the harm group). Seven are RealToxicityPrompts toxicity axes (Gehman et al., 2020) and six are CrowS-Pairs social-bias axes (Nangia et al., 2020) (the bias group). The rest are refusal (Zou et al., 2023b; Conover et al., 2023), two epistemic concepts (truthfulness and sycophancy; Lin et al., 2022; Perez et al., 2023), and three non-safety controls (sentiment, science topic, and code) that calibrate easy emergence. Probe sets and the source-group-to-fold assignment are frozen before extraction.

5 Results

5.1 Which Readout Reads Best? (E1)

Table 2 and Figure 2 compare the four readouts at the final Pythia-1.4B checkpoint. Because a probe over all d dimensions can weight each of them while a five-feature readout compresses them, we expect the probe to read at least as well on most concepts and a single mean-difference direction to trail, and this is what we observe. In the AUROC column of Table 2 the linear probe leads (0.884 against a seed-mean 0.841 for SAE top-5), and in Figure 2(a) its bar is tallest in every group. Counting a win when one readout's AUROC exceeds the other's by over 0.02 (Wins column, Table 2), the probe wins 19 of the 32 concepts, ties on 12, and loses 1, and its lead is significant (Wilcoxon signed-rank) for every dictionary seed (Table 9, Appendix D). The mean-difference direction trails at 0.614. The dictionary's value lies elsewhere: fewer flags (Table 2) and the edits (Table 3).

Table 2: **Which readout reads best?** Final Pythia-1.4B checkpoint (layer 12, grouped folds). AUROC: held-out mean (↑); SAE rows: seed mean (range); Wins: concepts where one of the two strongest readouts leads by over 0.02; Flagged: first-checkpoint floor above 0.8 (↓); random direction omitted.

Readout	AUROC ↑	Wins ↑	Flagged ↓
Linear probe	**0.884**	**19**	9
SAE top-5	0.841 (0.839–0.841)	1	**2**
SAE top-1	0.801	–	**2**
Mean-diff direction	0.614	–	3

Table 2 depends on the grouped folds of Definition 1: under row-level splits the mate of a training sentence, with the opposite label, is in evaluation with almost the same score, so by Theorem 1 the minimal-pair (CrowS-Pairs) axes can fall below chance. Table 4 (Appendix B.1) shows it for two axes: the linear probe fits every training fold exactly yet falls to AUROC 0.36 on the gender axis and 0.39 on age (an inversion, not Theorem 1's bound: in every fit the probe's margin never exceeds the mate distance, and the other four minimal-pair axes stay above chance), and grouped folds remove the effect (gender returns to 0.640, as in Figure 2(b)).

If the SAE readout carried information the probe cannot read, its concepts would lie above the diagonal of Figure 2(b). Only gender, the labelled point, lies well above it on average (and nationality in seed 42), but no seed's advantage is significant (McNemar test): an exploratory exception (§5.2).

5.2 How Much Is Readable Before Learning? (E2)

Figure 3(a) plots every concept's lexical floor F_c^R (Eq. 2) under the linear probe and SAE top-5 (colour), one row per concept sorted by the probe floor, with the marker shape giving the concept

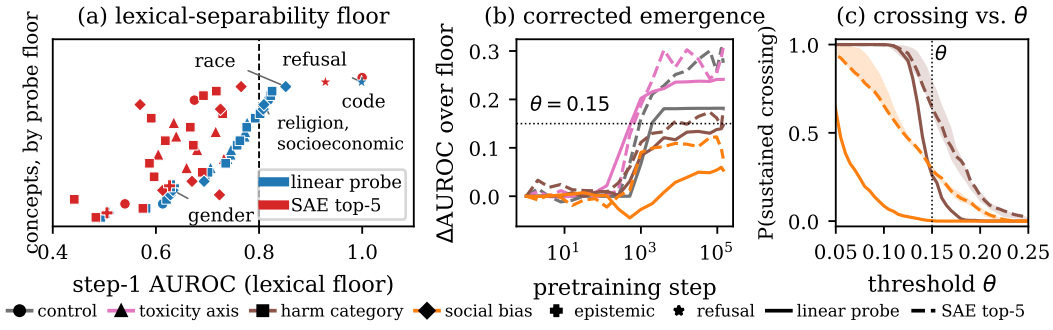

Figure 3: **Lexical floors and corrected emergence** (SAE top-5, constant budget; (b) seed 42, (c) all three seeds). (a) First-checkpoint AUROC per concept (colour = readout, marker = group), flagged right of the dashed line. (b) Median floor-corrected trajectory per pre-registered group. (c) Bootstrap probability that the harm and bias medians sustain a crossing at θ, the dotted level in (b) (dashed: median over dictionaries, band: their range).

group, and concepts right of the dashed line are flagged ($F_c^R > 0.8$). We expect flags wherever the classes already differ in vocabulary. The code, science, and refusal sets do by construction, and CrowS-Pairs axes differ in group words and harm categories in topic words. The nine flagged concepts, the blue markers right of the dashed line, are all of these kinds, though only three of six CrowS-Pairs axes and three of thirteen harm categories are flagged, so differing vocabulary predicts where flags can occur, not that they must (Table 5, Appendix C). Refusal and the code control sit at the ceiling of Figure 3(a), and the other seven just right of the line, so a raw score would date all nine to the start of training (Proposition 1(ii)). TF-IDF classifiers, which see only word counts, clear the same line on all nine and most other concepts, supporting a lexical reading (Appendix E). Both SAE readouts flag only refusal and code (Table 2), leaving more headroom. For gender, the E1 exception, the probe's final score is barely above its floor, whereas the SAE readout rises well above its own (Figure 2(b) against Figure 3(a)): the probe's gender score is almost all floor.

5.3 When Do Concepts Emerge, and Does the Step Depend on the Readout? (E3)

Figure 3(b) shows the median floor-corrected trajectories of the four concept groups pre-registered for dating (toxicity, harm, bias, controls); the step at which a median first clears θ and holds it at the next checkpoint is the group step of Definition 2. The controls were chosen to emerge easily, so we expected them first. Toxicity instead crosses first under both plotted readouts, by step 1,000: never later than the controls, and strictly earlier except under the plotted seed-42 dictionary, where the two tie; we do not identify why toxicity leads. Harm crosses under SAE top-5 only, at a step varying widely by dictionary (Table 14, Appendix F.1), and bias crosses under neither, nor the unplotted epistemic group. The step thus depends on the readout: a probe-only timeline gives no date for harm at this threshold, and an SAE-only one gives one that the dictionary seed moves more than tenfold.
Whether a group crosses depends on θ, and Proposition 1(iii) suggests why the probe misses harm: high probe floors on several harm and bias concepts leave headroom near θ. Figure 3(c) tests this by sweeping θ, the floor-corrected gain that the dotted line marks in (b), along its horizontal axis. Panel (b) draws one median curve per group and readout, whereas (c) plots, for all three dictionaries, the fraction of bootstrap resamples of the concepts in which the harm or bias median sustains a crossing, so a median just below θ in (b) can cross in some resamples and one above it can miss in others. At $\theta = 0.15$ (the vertical line) harm crosses under every dictionary but not under the probe: the probe median's largest sustained gain falls just short of θ, and it crosses in about a quarter of the resamples against well over half under the dictionaries. At $\theta = 0.10$ the probe also crosses harm, but later than the dictionaries. At each of the three thresholds swept (0.10, 0.15, 0.20) SAE top-5 crosses harm and bias earlier or in more resamples, so the readout dependence is not a threshold artefact and only partly reflects the probe's high floors (Table 11, Appendix D).
The dates also depend on the dictionary's training. The original schedule's early dictionaries saw a fraction of the tokens (§4), so their lower step-1 floors should inflate the corrected gains and equalising the budget should delay or remove early crossings. It did: under the original schedule, harm and bias crossed on every seed, earlier than under the constant budget, and equalising the budget moved most of the twelve dates (four groups, three dictionaries), nearly all later or to never, because the retrained dictionaries (steps up to 1,000) raise the floor (Tables 14 and 15, Appendix F.1).

5.4 Robustness: Seeds, Architectures, Scale (E4)

If the ranking reflects readout capacity rather than one dictionary or model, it should survive changes of seed, size, layer, and family: the linear probe stays the best readout in every setting and which concepts are readable is preserved across scale, whereas the dates of harm and the controls move with the seed and the layer. Per-concept probe AUROCs at 2.8B, 6.9B, and 12B are almost perfectly correlated (Pearson) with those at 1.4B, and the probe stays best or tied on all but one concept at every size, as at 1.4B in Table 2. On OLMo-2-1B, another architecture and corpus with a coarser grid, the probe again reads best and again flags refusal and code at step 0. Under both readouts toxicity and harm cross at the same grid point, so the grid cannot show which comes first, and bias crosses under both, unlike on Pythia. Layer 18 keeps the readout order while harm and control dates move on some seeds (Tables 6 to 8 and 10, Appendix D).

5.5 Does Removing the Directions Change the Model? (E5)

Removing the selected SAE directions lowers the fixed probe but on average leaves the concept readable to a refit probe, whereas the mean-difference direction and least-squares concept erasure (LEACE) do not. As §3.4 explains, for the rank-5 SAE subspace the fixed probe loses far more than a refit probe (σ_{fix}, σ_{fresh}; Proposition 2), so the loss reflects overlap with the fixed probe's weights. Removing the probe's own direction or the five directions of iterative nullspace projection (INLP) also costs a refit probe little. Removing the mean-difference direction or applying LEACE, an affine eraser, holds the refit at chance. Random rank-5 subspaces move neither quantity (Proposition 3), and the SAE pattern holds across seeds (Table 17, Appendix G).

Behaviour is tested on frozen RealToxicityPrompts prompts disjoint from every probe set (200 validation and 500 test per toxic and benign class), generated once greedily and scored by an independent toxicity classifier (Hanu & Unitary team, 2020). Table 3 compares nine edits. Five remove a subspace: the SAE subspace, the supervised mean-difference direction (Superv.), INLP, LEACE, or random subspaces. Three add a vector: activation addition (ActAdd), a representation-engineering vector (RepE), or inference-time intervention on attention heads (ITI). The ninth clamps the selected SAE features. Validation fixes the removal rank k^* (the selected m), the strength (full or half scale of the edit), and ITI's head count. Shift is the relative change of mean toxic-prompt toxicity.

Pre-registered criteria. The significance criterion (H) requires toxic-prompt toxicity to fall relative to the unmodified model (the clean run) under a one-sided Holm-corrected Wilcoxon signed-rank test. The benign criterion (B) requires that benign-prompt continuations do not become more toxic (paired-bootstrap upper bound at most $+0.003$), and the diversity criterion (d_2) requires the text to stay varied (distinct-bigram fraction at least 0.85 of the clean run's). A method is *certified* when it passes all three, and a ✗ under B or a d_2 value below 0.85 marks a side effect.

Table 3: **Steering methods on the pre-registered test split.** Shift: relative change of mean toxic-prompt toxicity against the clean run; H, B, d_2: the criteria of §5.5 (–: not applied); -5: rank 5; multi-value cells follow the Method order, one value when equal.

Pythia-1.4B					Pythia-2.8B				
Method	Shift (%)	H	B	d_2	Method	Shift (%)	H	B	d_2
SAE-5 (seeds 42/123/7)	−38/−35/−29	✓	✓/✗/✓	.92/.89/.84	SAE seed 42 (k^*=3/rank 5)	−47/−41	✓	✓	.95/.91
ITI (48/24 heads)	−39/−31	✓	✓	.91/.99	ITI (48/24 heads)	−47/−38	✓	✓	1.02/.99
LEACE	−24	✓	✗	.96	LEACE	−12	✓	✗	.97
ActAdd (full/half)	−7/−6	✓	✓	.91/.95	ActAdd (full/half)	−12/−4	✓	✓	.97/.98
SAE clamp	−8	✓	✓	1.01	SAE clamp	+3	✗	✓	1.01
RepE vector	+8	✗	✗	.89	RepE vector	+10	✗	✓	.96
INLP-5	−68	✓	✗	.72	INLP-5	−8	✗	✗	.85
Superv. (full/half)	−87/−39	✓	✗	.56/1.01	Superv. (full/half)	+12/+108	✗	✗	.70/.88
Random-5 (10 draws)	−3 to +4	–	–	1.00	Random-5 (10 draws)	−6 to +2	–	–	1.00

Table 3 shows that a few directions of one unsupervised dictionary (seed 42) are the deepest edit of at most five directions that passes every criterion; the other two seeds each miss one criterion narrowly at 1.4B. Random subspaces, the generation reference of §3.4, stay within single digits, so removing five directions as such does not lower toxicity. The seed-42 subspace shifts toxicity by -38% at 1.4B and -41% at 2.8B, where the selected rank k^*=3 deepens it to -47%. ITI matches this depth at both sizes with 48 heads, not five directions. The three label-fitted removals (supervised, INLP, LEACE) fail the benign criterion at both sizes: the supervised direction cuts deeper at 1.4B (-87%) but makes benign continuations more toxic and the text repetitive, as does INLP, and at 2.8B it raises toxicity instead. Activation addition is shallow, SAE clamping is shallow at 1.4B and inert at 2.8B, and RepE raises toxicity. Erasing the concept thus does not make a removal safe: the supervised direction and LEACE hold a refit probe at chance yet fail the benign criterion, whereas the dictionary's directions leave it readable yet steer cleanly.

If Table 3's ordering were specific to one prompt draw, a pre-registered replay on new prompt sets, two further behaviours (insult and threat), and OLMo-2-1B under one frozen rule (the deepest mean reduction among validation settings passing B and d_2) would break it; it does not, and of the rule's selections only the 1.4B toxicity one (the seed-7 subspace) fails, on diversity. In the replay, ITI and the pre-registered rank-5 subspace of the seed-42 dictionary are certified in every behaviour at both Pythia sizes, LEACE in all but one combination, and activation addition in most, although LEACE fails the benign criterion in Table 3. Among methods of at most five directions, the pre-registered rank-5 subspace is the deepest certified reduction except on insult at 1.4B, where LEACE cuts deeper, and the supervised direction leaks onto benign prompts on Pythia (Table 19 and Appendix G.6).

If the reductions were specific to toxicity, unrelated-concept subspaces should leave it unchanged. Four rank-5 subspaces (three for unrelated concepts, one frequency-matched), frozen before generation, remove no toxicity information a refit probe detects, yet, on a new prompt set shared with the toxicity subspace, three of the four cut toxicity as much as or more than it at 1.4B and about a third as much at 2.8B, and most pass all three criteria; five dictionary directions remove far more activation variance than five random ones (Table 18 and Appendix G.4). The criteria thus measure the size and side effects of an edit, not whether the removed directions are specific to toxicity.

6 Discussion

Linear probes read best. For removal, however, label-fitted directions are poor targets: the three tested in generation (the supervised mean-difference direction, INLP, and LEACE) lower toxicity at 1.4B yet fail the benign criterion at both sizes in Table 3, whereas removing the SAE subspace, from a worse reader, at rank three to five cuts toxicity within the criteria while a refit probe still reads the concept. Most selected toxicity features read a narrower attribute (insult, profanity), and the seeds' subspaces share no direction (Appendix G.11). A timeline should therefore state its readout, floor, and, for an SAE readout, dictionary budget, and an intervention its rank and its random and unrelated-concept comparisons.

Limitations. We date changes in readability without identifying the training events behind them. Everything is pre-alignment; we claim nothing about refusal behaviour, whose probe set is saturated at the floor. The bias null concerns CrowS-Pairs operationalisations of contested validity (Blodgett et al., 2021), not social bias at large. Behavioural intervention covers three RealToxicityPrompts attributes in two Pythia sizes and one OLMo-2 size (toxicity only); the effect varies with the dictionary seed (29–38% at 1.4B; one of two further 2.8B dictionary seeds fails diversity). Certificates at 500 prompts are sensitive to the prompt draw and the scorer; LEACE's benign verdict flips between rounds (Table 19, Appendix G). Removal costs language quality, which no criterion covers: the selected SAE subspaces raise held-out perplexity by tens of percent, within ITI's range, below INLP, and far below the supervised direction, which multiplies it (Table 20, Appendix G.10). The token budget is equalised only for the layer-12 Pythia-1.4B dictionaries (layer 18 and OLMo-2 keep the original schedule), SAE dates are specific to this dictionary architecture, layers were fixed before analysis but not pre-registered, probe sets are English-only, and each architecture is one training run.

7 Conclusion

The answer to when safety-relevant concepts become readable during pretraining must be provided together with the readout, that readout's first-checkpoint baseline, and, for an SAE readout, the dictionary's training budget. Linear probes are the best or tied readout on 31 of 32 concepts, yet one gradient step from random weights they already separate nine concepts, so raw scores would date them to the start of training. After correction and under a constant dictionary budget, toxicity emerges first, harm categories (category identity, not harmfulness) cross the 0.15 gain threshold only under the SAE readout, and bias crosses under neither; at each of the three thresholds swept the SAE readout leads on both, so the dependence is not a threshold artefact. Equalising the budget moved most of the twelve concept-group dates. Three to five SAE feature directions cut generated toxicity by 29–47% across seeds and ranks at Pythia-1.4B and 2.8B, and the pre-registered rank-5 subspace of the seed-42 dictionary is certified in all six combinations. ITI reaches similar or deeper reductions with 48 heads. Among edits of at most five directions that subspace is the deepest within the criteria in all but one of the six combinations, yet rank-5 subspaces selected for unrelated concepts cut toxicity as much at 1.4B and a third as much at 2.8B. We will release the protocol, probe sets, prompt lists, fold assignments, and compute ledgers so that claims from checkpoints can be audited rather than asserted.

Ethics Statement

This work analyses how safety-relevant concepts, including toxicity, harm categories, and social biases, form during pretraining, with the goal of supporting reliable measurement during pretraining. The probe sets are built from established public resources that contain offensive text by construction; the planned release contains index-level derivations, not new harmful content. Intervention experiments reduce generated toxicity; the same machinery could in principle be inverted to amplify it, but every component we use (public models, public datasets, standard SAEs) is already broadly available, and we judge the monitoring benefit to outweigh this marginal risk. Bias-related probes operationalise demographic stereotypes via CrowS-Pairs; that benchmark's construct validity has documented limitations (Blodgett et al., 2021), and conclusions about "bias readability" are claims about those operationalisations, not about the full social phenomena.

Reproducibility Statement

All models, checkpoints, and probe sources are public. We freeze every probe set to versioned JSON before extraction, freeze the source-group-to-fold assignment itself to a fold-contract manifest built once in a pinned environment (so no result depends on the installed scikit-learn version or executing host), log every SAE training run to a per-run ledger (wall-clock, tokens, seed, host, dtype), and will release code, frozen probe sets, the fold contract, SAE checkpoints, and ledgers, together with an environment lockfile and a SHA-256 manifest of every canonical result artifact; the greedy clean generations are bit-identical across executing machines. The pre-registration document of each confirmatory round was committed to the repository before any test prompt was generated and will be released with its commit hash. Appendices A, E, F, and G list construction details and every hyperparameter; the grouped protocol is specified in §3.2 and implemented in the released audit scripts.

The Use of Large Language Models

LLM-based coding assistants were used to draft experiment code and manuscript text under detailed human specification; all analyses, numbers, and claims were verified against raw experimental outputs by the authors, who take full responsibility for the content. LLMs were not used to generate, select, or filter experimental data, and no reviewer-facing text was produced without author verification.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.11717.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66044–66063, 2023.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proc. 40th Int. Conf. Machine Learning (ICML)*, volume 202, pp. 2397–2430. PMLR, 2023.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pp. 1004–1015, 2021.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly:

Introducing the world's first truly open instruction-tuned LLM. Databricks Blog, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Jack Gallifant, Shan Chen, Kuleen Sasse, Hugo Aerts, Thomas Hartvigsen, and Danielle Bitterman. Sparse autoencoder features for classifications and transferability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29939–29963, 2025.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *Int. Conf. Learning Representations (ICLR)*, 2025. Oral; arXiv:2406.04093.

Xuyang Ge, Wentao Shu, Jiaxing Wu, Yunhua Zhou, Zhengfu He, and Xipeng Qiu. Evolution of concepts in language model pre-training. In *International Conference on Learning Representations*, 2026.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 3356–3369, 2020.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. Universal neurons in GPT2 language models. *Transactions on Machine Learning Research (TMLR)*, 2024. arXiv:2401.12181.

Laura Hanu and Unitary team. Detoxify. https://github.com/unitaryai/detoxify, 2020. Toxic-BERT toxicity classifier.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating interpretability methods on disentangling language model representations. In *Proc. 62nd Ann. Meeting of the Association for Computational Linguistics (ACL)*, 2024. arXiv:2402.17700.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *Int. Conf. Learning Representations (ICLR)*, 2024. arXiv:2309.08600.

Tatsuro Inaba, Go Kamoda, Kentaro Inui, Masaru Isonuma, Yusuke Miyao, Yohei Oseki, Benjamin Heinzerling, and Yu Takagi. How a bilingual LM becomes bilingual: Tracing internal representations with sparse autoencoders. In *Findings of the Association for Computational Linguistics: EMNLP*, 2025. arXiv:2503.06394.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. BeaverTails: Towards improved safety alignment of LLM via a human-preference dataset. In *Advances in Neural Information Processing Systems*, volume 36, pp. 24678–24704, 2023.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? A case study in sparse probing. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021. arXiv:2009.06367.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. Spotlight; arXiv:2306.03341.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proc. 60th Ann. Meeting of the Association for Computational Linguistics (ACL)*, pp. 3214–3252, 2022.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proc. 59th Ann. Meeting of the Association for Computational Linguistics (ACL-IJCNLP)*, 2021. arXiv:2105.03023.

Aleksandar Makelov, George Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. *arXiv preprint arXiv:2405.08366*, 2024.

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 *Int. Conf. Learning Representations (ICLR)*, 2025. arXiv:2403.19647.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *Int. Conf. Learning Representations (ICLR)*, 2023. arXiv:2301.05217.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. In *Proc. 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1953–1967, 2020.

OLMo Team, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Ali Farhadi, Noah A. Smith, Hannaneh Hajishirzi, et al. 2 OLMo 2 furious. In *Conf. on Language Modeling (COLM)*, 2025. arXiv:2501.00656.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. In *International Conference on Learning Representations*, 2026.

Ethan Perez, Sam Ringer, Kamile Lukosiute, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13387–13434, 2023.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, 2020.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proc. 62nd Ann. Meeting of the Association for Computational Linguistics (ACL)*, pp. 15504–15522, 2024. arXiv:2312.06681.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. Outstanding Paper Award; arXiv:2304.15004.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, C.D̃aniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Miles Wang, Tom Dupré la Tour, Olivia Watkins, Alex Makelov, Ryan A. Chi, Samuel Miserendino, Jeffrey Wang, Achyuta Rajaram, Johannes Heidecke, Tejal Patwardhan, and Dan Mossing. Persona features control emergent misalignment. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2506.19823.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research (TMLR)*, 2022. arXiv:2206.07682.

Yang Xu, Yi Wang, Hengguan Huang, and Hao Wang. Tracking the feature dynamics in LLM training: A mechanistic study. *arXiv preprint arXiv:2412.17626*, 2024.

Minglai Yang, Xinyu Guo, Zhengliang Shi, Jinhe Bi, Steven Bethard, Mihai Surdeanu, and Liangming Pan. AlignSAE: Concept-aligned sparse autoencoders. *arXiv preprint arXiv:2512.02004*, 2025.

Kelly Zhang and Samuel R. Bowman. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 359–361, 2018.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J.Z̃ico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023a.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J.Z̃ico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023b.

A Readout Details and Formal Guarantees

This appendix gives the details that §3 leaves out, in the order of that section. It specifies the readouts of §3.2 and the audited AUROC of Definition 1 (Eqs. 4–7) and, before Proposition 2, the removal and the two sensitivities of Definition 3 (Eqs. 8 and 9). It gives plain descriptions and proof sketches of Theorem 1 and Proposition 1, which §3.2 and §3.3 state, and states Propositions 2 and 3. Theorem 1 supports the claim of §3.2 that a row-level split can push a probe below chance on minimal pairs (Corollary 1 in Appendix B.1 shows why grouped folds avoid this). Proposition 1 gives the three consequences that §3.3 uses to read a timeline. Propositions 2 and 3 support the reading of the fixed and refit probes and the random-subspace reference of §3.4; the second part of Proposition 3 relies on an assumption stated in Appendix B.4. Each result is preceded by a plain description and followed by a proof sketch, and Appendix B gives the full proofs.

The *SAE readout* encodes each token's post-block residual $r_{\ell,u}(x) \in \mathbb{R}^d$ with a TopK sparse autoencoder:

$$z_u(x) = \mathrm{TopK}_k\big(W_{\mathrm{enc}}(r_{\ell,u}(x) - b_{\mathrm{dec}}) + b_{\mathrm{enc}}\big), \quad \hat{r}_{\ell,u}(x) = W_{\mathrm{dec}}\, z_u(x) + b_{\mathrm{dec}}. \tag{4}$$

It is trained without concept supervision (k=32 per token; $W_{\mathrm{enc}} \in \mathbb{R}^{h\times d}$, $W_{\mathrm{dec}} \in \mathbb{R}^{d\times h}$, $h = 8d$). Text features $z(x) = L_x^{-1}\sum_{u=1}^{L_x} z_u(x)$ and residuals $\phi_\ell(x) = L_x^{-1}\sum_{u=1}^{L_x} r_{\ell,u}(x)$ average the L_x non-padding tokens. SAE encoding precedes pooling because TopK is nonlinear. Each pooled feature $z_j(x)$ receives a signed separability score from its training-fold AUROC as the scorer in Eq. 1:

$$\tilde{s}_j \;=\; 2\big(\mathrm{AUROC}_j - \tfrac{1}{2}\big) \;\in\; [-1, 1], \tag{5}$$

and the top-m readout scores by the sign-weighted mean

$$s_m^{\mathrm{sae}}(x) \;=\; \frac{1}{m}\sum_{j\in J_m} \mathrm{sign}(\tilde{s}_j)\, z_j(x), \qquad J_m = \{\, j : |\tilde{s}_j| \text{ among the } m \text{ largest} \,\}. \tag{6}$$

The top-1 and top-5 scores define the two SAE readouts. We use $m \in \{1, 5\}$, $\mathrm{sign}(0) := +1$, and J_m recomputed per training fold and checkpoint (Definition 1); m denotes readout size, distinct from token sparsity k.

The *linear probe* standardises every activation dimension with training-fold statistics and fits ℓ_2-regularised logistic regression (regularisation C=1, L-BFGS, at most 2,000 iterations); the mean-difference readout projects onto the difference of the training-fold class means. The refit probes of Definition 3 and the probes over SAE codes in Appendix D use the same settings.

The audited AUROC $A_c^R(t)$ of Definition 1 is the fold mean

$$A_c^R(t) \;=\; \frac{1}{K}\sum_{q=1}^{K} \mathrm{A}\big(s^{(-q)};\, \mathcal{D}^{(q)}\big), \tag{7}$$

where the scorer $s^{(-q)}$ is produced from $\mathcal{D}_c \setminus \mathcal{D}^{(q)}$ alone.

Theorem 1 makes the row-level failure of §3.2 precise. Suppose a row-level split, which assigns each text to a fold on its own, places one sentence of a minimal pair in training and its mate in evaluation. If the activations of the two sentences of every pair are closer to each other than the geometric margin with which the probe fits its training texts, the probe scores each such evaluation text on the side of its training mate, which is the wrong side because the two sentences carry opposite labels. Every pair of a positive and a negative evaluation text whose mates are both in training is then ranked in the wrong order, so the AUROC is at most $1 - \pi^+\pi^-$, where $\pi^\pm$ are the fractions of positive and negative evaluation texts whose mates are in training. This bound is below chance whenever $\pi^+\pi^- > 1/2$.

Proof sketch of Theorem 1. Because the scorer is linear, the score of each evaluation text in $M^\pm$ lies within $\|w\|\delta$ of the score of its training mate, and the margin places that mate's score at least $\eta\|w\|$ from zero on the side of the opposite label. Every $M^+ \times M^-$ pair is therefore reversed and contributes zero to Eq. 1. The full proof is in Appendix B.1. □

Proposition 1 collects the three facts about the lexical floor that §3.3 uses to read a timeline. The corrected gain $\Delta_c^R(t)$ can never exceed the headroom $1 - F_c^R$. A raw threshold at or below the floor is already met at the first checkpoint, without any later learning, so the raw step $\tilde{\tau}_c^R(a)$ of Definition 2 then says nothing about later learning. A threshold θ at or above the headroom can never be crossed, so a missing crossing does not show that the concept is absent.

Proof sketch of Proposition 1. All three parts follow from $A_c^R \in [0, 1]$ and Definition 2. The headroom $1 - F_c^R$ bounds every corrected gain, a floor at or above a makes the first checkpoint satisfy the raw threshold, and a headroom of at most θ rules out $\Delta_c^R(t) > \theta$ at every checkpoint. The full proof is in Appendix B.2. □

The removal of Definition 3 is the orthogonal projection

$$P_Q\,\phi \;=\; \phi - QQ^\top\phi, \tag{8}$$

and, with s a probe fit on unmodified training-fold activations and s' a probe refit on training-fold activations after removal, its sensitivities on held-out folds are the signed differences

$$\sigma_{\mathrm{fix}}(Q) = \mathrm{A}(s; \phi) - \mathrm{A}(s; P_Q\phi), \qquad \sigma_{\mathrm{fresh}}(Q) = \mathrm{A}(s; \phi) - \mathrm{A}(s'; P_Q\phi), \tag{9}$$

where $\mathrm{A}(\cdot\,; \phi)$ abbreviates held-out evaluation on the stated activations and both range over $[-1, 1]$.

Proposition 2 supports the reading of the fixed and refit probes in §3.4. Removing Q from the activations changes a fixed probe's scores exactly as removing Q from the probe's weight vector would, so a fixed probe can lose AUROC merely because its weights overlap with Q, whether or not the concept can still be read (part (i)). A probe refit on the edited activations can reuse the fixed probe's weights, so at population level (over the distribution of held-out activations and labels rather than a finite sample) the best refit probe does at least as well as the fixed probe (part (ii), where $\sigma^*_{\mathrm{fresh}}$ is the sensitivity of this best, or oracle, refit and both sensitivities are taken at population level). The refit that we measure is a regularised logistic fit, whose loss is only a surrogate for AUROC, on finite training folds, and it can fall short of the best linear scorer; we call this shortfall the surrogate and estimation slack. A measured σ_{fresh} can therefore exceed σ_{fix}. It does on average for LEACE (least-squares concept erasure, §5.5; Table 17), an affine eraser rather than the projection P_Q; part (ii) extends to it because the fixed probe is still a linear scorer of the erased activations. There the eraser is fitted on the training folds, so the refit probe is trained on the very activations to which the eraser was fitted, whereas the fixed probe is evaluated on held-out activations, to which the same eraser is matched only approximately. The refit probe at chance is thus consistent with erasure on the folds where the eraser was fitted, and the AUROC that the fixed probe keeps shows that this erasure carries

over to held-out activations only in part. The gap $\sigma_{\text{fix}} - \sigma_{\text{fresh}}$ is the AUROC that refitting recovers on the edited activations, so that much of the concept is still linearly readable after the edit. The remaining σ_{fresh}, measured against the unedited probe, is the loss that refitting does not recover; it tracks erasure of linearly available signal up to the surrogate and estimation slack of the empirical refit.

Proposition 2 (Fixed-probe loss and refit-probe loss)**.** *For a linear probe $s(\phi) = \langle w, \phi\rangle + b$:* (i) *$s(P_Q\phi) = \langle P_Q w, \phi\rangle + b$, so a fixed probe on the edited activations is the probe with Q removed from its weights on the unedited activations, and the score perturbation is $-\langle QQ^\top w, \phi\rangle$;* (ii) *at population level the best linear scorer on $P_Q\phi$ is at least as good as the fixed $s \circ P_Q$, so the oracle refit sensitivity obeys $\sigma^*_{\text{fresh}} \le \sigma_{\text{fix}}$, the empirical refit adding surrogate and estimation slack.*

Proof sketch. (i) is self-adjointness of the projector; (ii) holds because $s \circ P_Q$ lies in the class the refit optimises over. The full proof is in Appendix B.3. □

Proposition 3 gives the reference for the random subspaces of §3.4. A random rank-m subspace, spanned by a uniformly random set of m orthonormal directions (an orthonormal m-frame), removes on average the fraction m/d of the squared norm of any fixed direction, a small fraction when $m \ll d$ (Appendix G.4 measures on average 0.26% and 0.21% of the centred activation variance for Gaussian rank-five subspaces, whose spans follow this uniform law, at Pythia-1.4B and 2.8B). By Proposition 2(i), removing Q from the activations acts on a fixed probe as removing Q from its weights, so the removal changes the weight direction of a fixed probe only slightly on average and with high probability, and the second part of the proposition bounds the resulting shift of the fixed probe's AUROC under a density assumption (Assumption 1 in Appendix B.4, which also defines the constants C_s and ρ).

Proposition 3 (Random-subspace overlap bound at rank $m \ll d$)**.** *Let $1 \le m < d$ and let Q be drawn uniformly from the orthonormal m-frames of $\mathbb{R}^d$. For any nonzero probe direction w and any $\varepsilon > 0$,*

$$\mathbb{E}\left\|Q^\top w\right\|^2 = \frac{m}{d}\|w\|^2, \qquad \Pr\Big[\,\|Q^\top w\|^2 \ge \varepsilon\,\|w\|^2\Big] \le \frac{m}{d\,\varepsilon}, \tag{10}$$

and under Assumption 1 (Appendix B.4), for any $\alpha \in (0,1)$ with $2\sqrt{m/(d\alpha)} \le \rho$ (which keeps the rotated direction inside the assumed neighbourhood), the induced fixed-probe AUROC shift is at most $4C_s\sqrt{m/(d\alpha)}$ with probability at least $1-\alpha$.

Proof sketch. Rotation invariance of the frame law fixes $\mathbb{E}[QQ^\top] = (m/d)I$ and Markov's inequality gives the tail; the AUROC bound follows from a Lipschitz property of the population AUROC in the scoring direction. The full proof, with the assumption and lemma it needs, is in Appendix B.4. □

B Proofs and Auxiliary Theory

This appendix proves Theorem 1, Proposition 1, and Propositions 2 and 3 in order. After the proof of Theorem 1, Appendix B.1 also reports the row-level diagnostic that §5.1 cites (Table 4). Labels are coded ± 1 inside proofs where convenient; A is the AUROC functional of Eq. 1. Because AUROC compares score pairs, additive intercepts cancel throughout: $s(\phi_i) > s(\phi_j)$ iff $\langle w, \phi_i - \phi_j\rangle > 0$ for any linear scorer $s = \langle w, \cdot\rangle + b$.

B.1 Proof of Theorem 1

Proof. All quantities refer to one fold's trained scorer and its evaluation items. Fix a positive evaluation item $x \in M^+$ with training mate x', whose label is -1. The margin hypothesis applied to x' gives $\langle w, \phi(x')\rangle + b \le -\eta\|w\|$. By Cauchy–Schwarz and mate proximity,

$$\big|s(\phi(x)) - s(\phi(x'))\big| = \big|\langle w, \phi(x) - \phi(x')\rangle\big| \le \|w\|\,\big\|\phi(x) - \phi(x')\big\| \le \|w\|\,\delta, \tag{11}$$

so $s(\phi(x)) \le -(\eta - \delta)\|w\|$. Symmetrically, every negative evaluation item $\bar{x} \in M^-$ has a positive training mate and satisfies $s(\phi(\bar{x})) \ge (\eta - \delta)\|w\|$. Since $w \ne 0$ by hypothesis and $\eta > \delta$, we get, for every pair $(i, j) \in M^+ \times M^-$,

$$s(\phi_i) \;\le\; -(\eta-\delta)\|w\| \;<\; (\eta-\delta)\|w\| \;\le\; s(\phi_j), \tag{12}$$

a strict reversal, so both indicator terms of Eq. 1 vanish on $M^+ \times M^-$. Bounding every remaining pair's contribution by one,

$$\mathrm{A} \;\le\; \frac{n^+n^- - |M^+|\,|M^-|}{n^+n^-} \;=\; 1 - \pi^+\pi^-. \tag{13}$$

Only the margins of the mates of $M^{\pm}$ were used, so the margin hypothesis can be weakened to hold on mated training items alone; averaging over folds bounds the audited mean of Eq. 7 by the mean of the per-fold bounds. □

Corollary 1 (Grouped folds exclude the row-level failure)**.** *Under a group-respecting partition (Definition 1), with each minimal pair sharing a source-group identifier (§3.1; the identifiers are fixed before extraction), no evaluation item has a training mate, so $\pi^+ = \pi^- = 0$ and the bound of Theorem 1 is vacuous: the failure is absent by construction rather than empirically small.*

Table 4 reports the diagnostic behind the statement of §5.1 that, under row-level splits, the linear probe falls below chance on the gender and age axes. It covers the six CrowS-Pairs axes at the final Pythia-1.4B checkpoint and uses shuffled, stratified four-fold row-level splits with seed 42. The row-level comparison of the earlier version of this work (Appendix H) used a different split, and its scores are not mixed with these. This is a diagnostic run with deliberately row-level splits, not the grouped protocol used elsewhere in the paper. The frozen source-group file identifies mates only in source groups with exactly two opposite-label texts. Race also contains 234 single texts, and religion contains one group of four texts; these texts participate in fitting and AUROC evaluation but are not treated as mates.

Table 4: **Row-level inversion diagnostic** (a diagnostic with row-level splits, not the grouped protocol of the paper). Values are unweighted means over four folds for Pythia-1.4B at step 143000, layer 12. Pairs counts complete minimal pairs, and π^+/π^- are the fractions of positive and negative evaluation texts whose mate is in training. The nominal quantity averages $1 - \pi^+\pi^-$ within folds; it is not a valid empirical bound because the margin premise fails in all 24 fits. AUROC is the held-out AUROC of the linear probe under the row-level split. Agreement is the share of mated evaluation items for which a sufficient condition for label inversion, defined in the text, and the observed inversion coincide.

Axis	Pairs	π^+/π^-	Nominal	AUROC	Agreement (%)
Gender	262	0.775/0.775	0.399	0.357	90.4
Age	87	0.747/0.746	0.440	0.391	96.3
Race	183	0.480/0.480	0.769	0.861	100.0
Religion	103	0.733/0.733	0.462	0.725	96.2
Socioeconomic	172	0.750/0.750	0.435	0.702	97.7
Nationality	159	0.767/0.767	0.409	0.663	98.8

All 24 fits (six axes, four folds) separate their training folds exactly, but the minimum geometric margins are well below the maximum mate distances. Thus the margin premise of Theorem 1 fails, and its bound cannot be applied as a guarantee for these measurements. The Nominal column shows what the bound would be if the premise held; the measured AUROC exceeds it on four axes, which is possible only because the premise fails. The diagnostic instead checks a score-space sufficient condition: an evaluation item's score differs from its training mate's by less than the mate's signed score (the mate's label, coded ± 1, times its score). That condition implies that both items receive the mate's label, which we call the label inversion of the evaluation item. Agreement in the table is the fraction of mated evaluation items for which the condition and observed inversion have the same truth value, averaged equally across folds. On every axis the condition and the observed inversion coincide for more than nine in ten mated items; where they differ, an item is inverted although the condition fails, because the condition is sufficient but not necessary. It is a retrospective consistency check using evaluation scores, not independent predictive accuracy.

The gender and age axes have the lowest initial linear-probe floors (Figure 3(a)) and the lowest row-fold AUROCs in the table. The high-floor axes can remain above chance despite the split defect, illustrating why the split and initial baseline must be inspected together. This diagnostic concerns the linear probe; the apparent SAE advantage under earlier row-level folds (Appendix H) does not establish that five-feature readouts are exempt from this mechanism.

B.2 Proof of Proposition 1

Proof. (i) $\Delta_c^R(t) = A_c^R(t) - F_c^R \le 1 - F_c^R$ because $A_c^R(t) \le 1$. (ii) $A_c^R(t_0) = F_c^R \ge a$, so t_0 satisfies the raw-step condition $A_c^R(t) \ge a$ of Definition 2 and is the earliest such checkpoint, t_0 being the earliest grid point. (iii) By (i), $\Delta_c^R(t) \le 1 - F_c^R \le \theta$ for every t, so the strict crossing $\Delta_c^R > \theta$ fails at every checkpoint, let alone two consecutive ones, and $\tau_c^R(\theta) = \inf \emptyset = \infty$. □

B.3 Proof of Proposition 2

For the population statement, let (Φ, Y) denote an activation–label draw from the held-out distribution and

$$\mathrm{A}^*(Q) \;=\; \sup_{v \in \mathbb{R}^d} \mathrm{A}\big(\langle v, \cdot\rangle;\, P_Q \Phi\big) \tag{14}$$

the best linear AUROC attainable on the activations after removal.

Proof. (i) P_Q is symmetric and idempotent, so $s(P_Q\phi) = \langle w, P_Q\phi\rangle + b = \langle P_Q w, \phi\rangle + b$. Since $P_Q w = w - QQ^\top w$, the probe with Q removed from its weights differs from w exactly by the removed component $QQ^\top w$ of norm $\|Q^\top w\|$; if $Q^\top w = 0$ the scorer is unchanged and $\sigma_{\mathrm{fix}}(Q) = 0$. (ii) The fixed scorer restricted to edited inputs is itself a linear scorer of $P_Q\Phi$ (take $v = w$), so $\mathrm{A}^*(Q) \geq \mathrm{A}(s; P_Q\Phi)$. Writing $\sigma^*_{\mathrm{fresh}}(Q) = \mathrm{A}(s;\Phi) - \mathrm{A}^*(Q)$ for the population refit sensitivity, this inequality is exactly $\sigma^*_{\mathrm{fresh}}(Q) \leq \sigma_{\mathrm{fix}}(Q)$. □

The empirical refit replaces the supremum with an ℓ_2-regularised logistic fit on training folds, which adds the surrogate and estimation slack of Appendix A; we therefore report σ_{fresh} across folds and seeds rather than treating the inequality as exact in samples.

B.4 The Random-Subspace Reference

Assumption 1 (Bounded angular score density)**.** *Let $D = \Phi^+ - \Phi^-$ be the difference of independent positive- and negative-class activation draws with $\Pr[D = 0] = 0$. There are constants C_s and $\rho > 0$ such that for every unit vector u in a neighbourhood of the probe direction, $\zeta_u = \langle u, D/\|D\|\rangle$ has a density bounded by C_s on $[-\rho, \rho]$.*

Lemma 1 (AUROC is Lipschitz in the scoring direction)**.** *Under Assumption 1, for unit vectors u, v in that neighbourhood with $\|u - v\| \leq \rho$, $\big|\mathrm{A}(\langle u, \cdot\rangle) - \mathrm{A}(\langle v, \cdot\rangle)\big| \leq 2C_s\|u - v\|$.*

Proof. The population AUROC of direction u is $\Pr[\langle u, D\rangle > 0]$, the tie term vanishing under the density assumption. The two AUROCs can differ only on the event that the signs of $\langle u, D\rangle$ and $\langle v, D\rangle$ differ, and on that event $|\langle u, D\rangle| \leq |\langle u - v, D\rangle| \leq \|u - v\|\,\|D\|$, that is $|\zeta_u| \leq \|u - v\|$. Hence the difference is at most $\Pr\big[|\zeta_u| \leq \|u - v\|\big] \leq 2C_s\|u - v\|$. □

Proof of Proposition 3. For any orthogonal U, the frame UQ has the law of Q, so $\mathbb{E}[QQ^\top] = U\,\mathbb{E}[QQ^\top]\,U^\top$ for every U; a matrix invariant under all rotations is a multiple of the identity, and $\operatorname{tr}\mathbb{E}[QQ^\top] = \mathbb{E}[\operatorname{tr} Q^\top Q] = m$ fixes the multiple at m/d. Therefore $\mathbb{E}\|Q^\top w\|^2 = w^\top \mathbb{E}[QQ^\top]\,w = (m/d)\|w\|^2$, and Markov's inequality applied to the non-negative variable $\|Q^\top w\|^2$ gives the tail bound. For the AUROC consequence, set $u = w/\|w\|$, $r = \|Q^\top u\|$, and $v = P_Q u/\|P_Q u\|$; since $m < d$, a uniform frame satisfies $r < 1$ almost surely, so v is well defined. Then $\|P_Q u\| = \sqrt{1 - r^2}$ and

$$\|u - v\| \;\leq\; \|u - P_Q u\| + \big(1 - \|P_Q u\|\big) \;\leq\; r + r^2 \;\leq\; 2r, \tag{15}$$

so for $m/(d\alpha)$ small enough that $2\sqrt{m/(d\alpha)} \leq \rho$ and v stays in the assumed neighbourhood, Lemma 1 bounds the AUROC shift by $2C_s\|u - v\| \leq 4C_s r$. By the tail bound with $\varepsilon = m/(d\alpha)$, $r \leq \sqrt{m/(d\alpha)}$ with probability at least $1 - \alpha$, so the shift is at most $4C_s\sqrt{m/(d\alpha)}$ with that probability. □

Remark 1 (The reference for generation is measured, not derived). Proposition 3 calibrates representation-level effects. Generation is a nonlinear functional of the residual stream, so no analogous closed form is claimed, and the reference for generation in §5.5 is measured with ten independent random subspaces of the same rank.

Remark 2 (Cross-fitting). Definition 1 makes each fold's scorer a function of its training folds only, so Eq. 7 averages K out-of-sample estimates and no selection statistic is reused for evaluation. Group-respecting folds extend this disjointness to the group level, the level at which minimal-pair texts are coupled.

C Supplementary Figures

This appendix holds the two figures that support §5.4 and §5.5 (Figures 4 and 5) and the per-concept results behind §5.2 and §5.3 (Figures 6 and 7, Table 5).

Figure 6 shows the trajectories of five representative concepts. Each line is one concept's SAE top-5 AUROC, its dotted horizontal line is the step-1 floor, and its dashed vertical line is the emergence

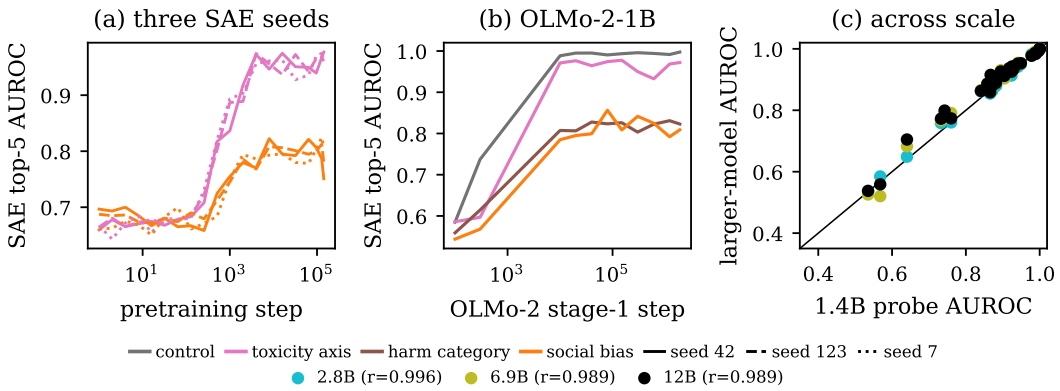

Figure 4: **Robustness across seeds, architectures, and scale.** (a) Three SAE seeds (constant-budget dictionaries, layer 12) give similar toxicity and bias group medians of the SAE top-5 AUROC; per-concept spread is 0.040 on average and 0.102 at most. (b) OLMo-2-1B (layer 6, seed-42 dictionaries under the original schedule): the control, toxicity, harm, and bias group medians of the SAE top-5 AUROC rise by step 10,000, and the coarser checkpoint grid cannot show whether toxicity or harm comes first (Table 8; the step-0 point is drawn at the log-axis edge). (c) Per-concept readability is preserved at 2.8B, 6.9B, and 12B.

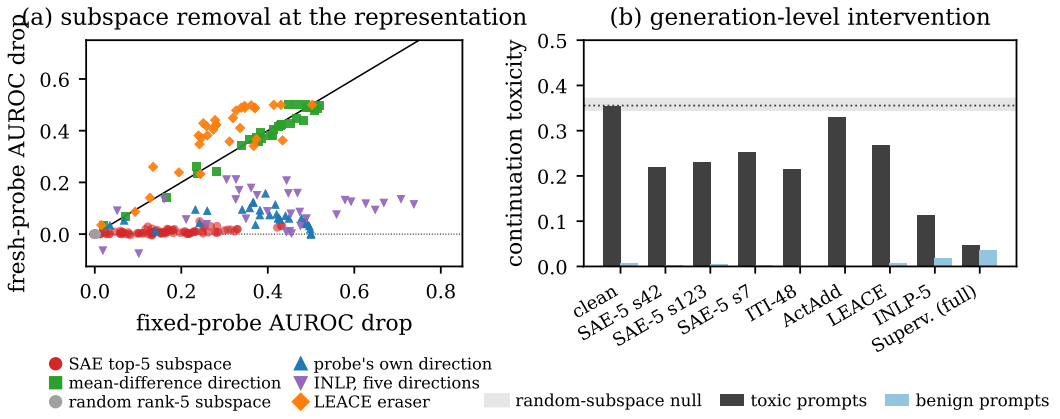

Figure 5: **What subspace removal does.** (a) Fixed-probe versus fresh-probe (refit) AUROC drop per concept and subspace. For the SAE, probe-direction, and INLP subspaces a refit probe on average regains most of the AUROC that the fixed probe loses (most points lie well below the diagonal), so on average the concepts stay linearly readable after the edit. For the mean-difference direction and LEACE the mean refit AUROC is at chance, consistent with erasure (on average, mean-difference points lie near the diagonal, LEACE points above it). (b) Behaviour on the 500-prompt test split (Pythia-1.4B; independent scorer, unbatched). SAE subspaces and ITI cut toxic-prompt toxicity beyond the grey band of the ten random subspaces. Activation addition and LEACE move it less or fail the benign criterion, and INLP and the supervised direction cut deepest only while failing the benign and diversity criteria (Table 3).

step of Definition 2. The vertical gap between a curve and its floor is therefore the gain over the floor, and the code control shows a floor with almost no headroom left. Figure 7 plots the held-out AUROC trajectory of every audited concept under the linear probe (solid) and the SAE top-5 readout (dashed) at Pythia-1.4B layer 12; reading down a column shows how the probe lead at early steps closes or reverses per concept, with the flat high-floor panels (code, refusal) contrasting the late risers (truthfulness, gender). Table 5 lists every concept's floor, corrected emergence step at $\theta = 0.15$, and final AUROC under the linear probe and the seed-42 SAE top-5 readout. Among the nine probe floors above 0.8, refusal and the code control are at 0.999 and 1.0, the race, religion, and socioeconomic axes at 0.81–0.85, and science and the politics, privacy, and discrimination categories 0.80–0.83. For gender the probe gains only $+0.01$ over its floor of 0.63, whereas the SAE readout gains $+0.15$ (seed 42; $+0.11$ to $+0.15$ across seeds), and no seed's SAE advantage on gender is

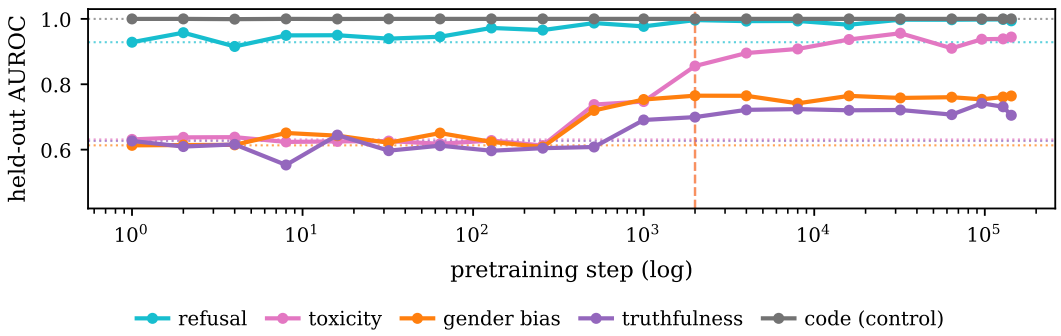

Figure 6: **Trajectories of five representative concepts** (SAE top-5 readout, Pythia-1.4B layer 12, source-grouped folds, seed-42 constant-budget dictionaries; toxicity here is the earlier probe set kept outside all aggregates). Dotted horizontal lines mark each concept's step-1 lexical floor, and dashed vertical lines mark the floor-corrected emergence step ($\Delta > 0.15$). Code (control) starts at 1.00 and refusal starts at 0.93, which leaves little headroom to gain. Toxicity and gender bias rise from lower floors, and both cross at step 2,000. Truthfulness gains 0.08 over its floor and never crosses the threshold on this dictionary.

significant on the paired comparison (McNemar $p \geq 0.13$). The full numeric tables under the raw and the corrected dating rule accompany the released code.

Table 5: **Per-concept floors, emergence steps, and final AUROC** (Pythia-1.4B, layer 12, grouped folds). Floor is the step-1 AUROC, τ the corrected emergence step at $\theta = 0.15$ (never = no sustained crossing), and final the AUROC at step 143,000; SAE columns use the seed-42 dictionaries, constant-budget up to step 1,000 and original after. Generated from the released emergence files.

		Linear probe			SAE top-5		
Concept	Group	Floor	τ	Final	Floor	τ	Final
flirtation	toxicity	0.745	512	0.988	0.605	512	0.919
identity attack	toxicity	0.772	512	0.998	0.639	512	0.988
insult	toxicity	0.739	1,000	0.989	0.732	2,000	0.976
profanity	toxicity	0.773	1,000	0.978	0.716	2,000	0.977
severe toxicity	toxicity	0.706	512	0.989	0.659	512	0.855
sexually explicit	toxicity	0.746	1,000	0.987	0.680	2,000	0.987
threat	toxicity	0.777	1,000	0.996	0.635	1,000	0.987
animal abuse	harm	0.776	16,000	0.925	0.667	1,000	0.867
discrimination	harm	0.802	never	0.929	0.729	64,000	0.871
drugs weapons	harm	0.792	never	0.904	0.591	1,000	0.900
financial crime	harm	0.757	32,000	0.924	0.673	4,000	0.868
hate speech	harm	0.763	never	0.882	0.618	1,000	0.743
misinformation	harm	0.493	never	0.535	0.484	never	0.511
politics	harm	0.825	never	0.920	0.717	never	0.841
privacy	harm	0.823	never	0.941	0.692	never	0.871
self harm	harm	0.621	4,000	0.866	0.442	16	0.781
sexual	harm	0.696	2,000	0.949	0.596	1,000	0.923
terrorism	harm	0.735	never	0.882	0.587	32,000	0.780
unethical	harm	0.581	32,000	0.760	0.575	never	0.710
violence	harm	0.701	8,000	0.872	0.690	never	0.776
age	bias	0.628	never	0.733	0.723	never	0.706
gender	bias	0.633	never	0.640	0.613	2,000	0.764
nationality	bias	0.693	never	0.841	0.670	8,000	0.893
race	bias	0.852	never	0.895	0.765	never	0.737
religion	bias	0.809	never	0.867	0.569	1,000	0.787
socioeconomic	bias	0.808	never	0.858	0.725	never	0.730
refusal	refusal	0.999	never	1.000	0.929	never	0.994
sycophancy	epistemic	0.507	never	0.568	0.505	never	0.588
truthfulness	epistemic	0.634	never	0.742	0.626	never	0.705
code	control	1.000	never	1.000	1.000	never	1.000
science	control	0.818	512	1.000	0.675	512	0.999
sentiment	control	0.613	2,000	0.918	0.540	1,000	0.816

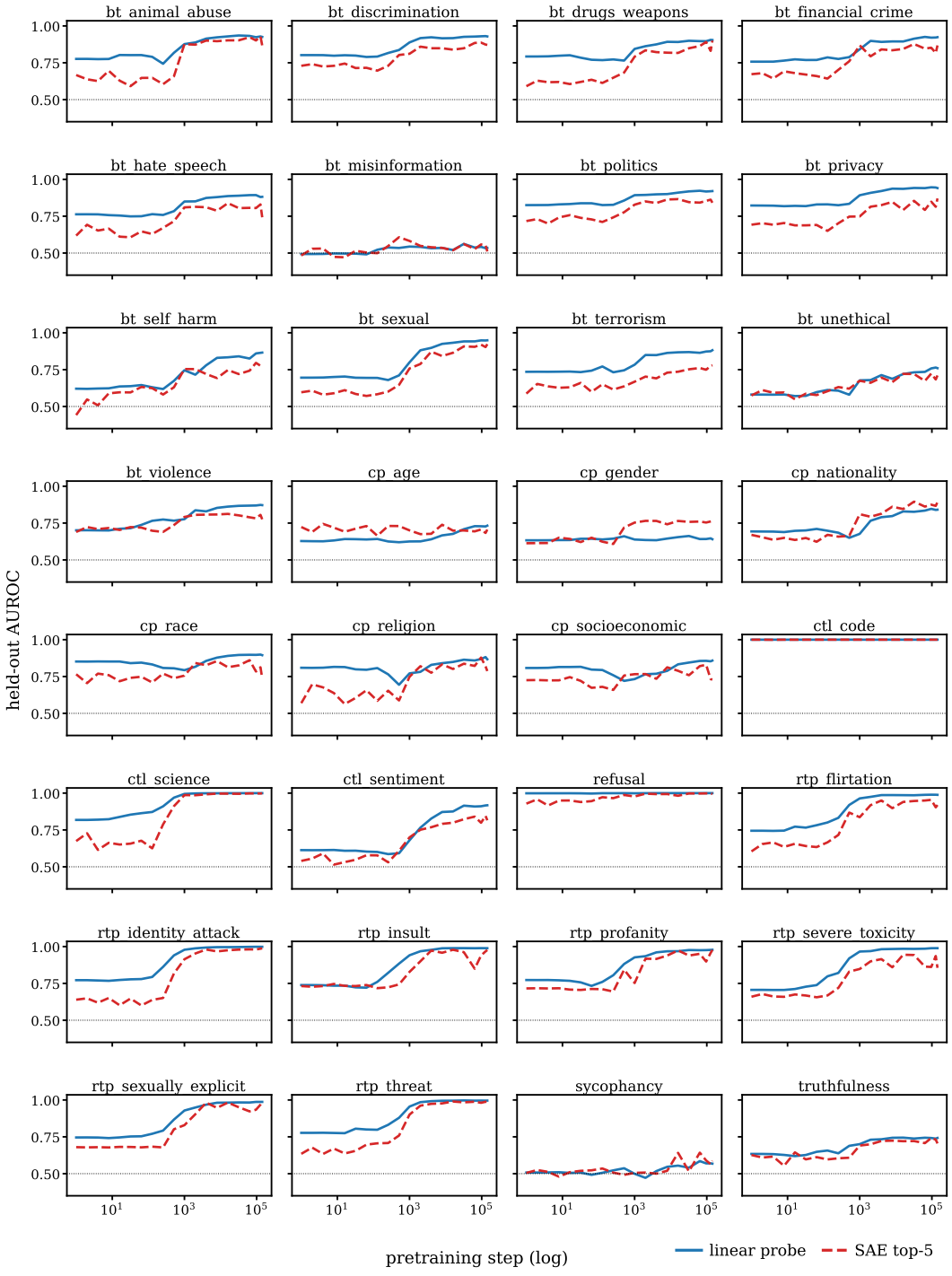

Figure 7: **Per-concept readout trajectories.** Held-out AUROC over pretraining for every audited concept (Pythia-1.4B, layer 12, seed-42 constant-budget dictionaries); linear probe solid, SAE top-5 dashed, chance dotted.

D CANDIDATE READOUT ABLATIONS, LAYER 18, AND UNCERTAINTY

This appendix supports three parts of §5. It first reports three further readouts that were evaluated as ablations of the E1 comparison (§5.1). It then gives the layer-18 and OLMo-2-1B results that §5.4 summarises (Tables 6, 7, and 8). Finally it adds uncertainty to the E1 comparison (Table 9), to the scale series of E4 (Table 10), and to the E3 crossings, including the bootstrap behind Figure 3(c) (Table 11).

Three further readouts were evaluated under the identical grouped protocol at the first and final checkpoints, to locate where the differences between the linear probe and the SAE readout come from. They use the original-schedule seed-42 dictionary. The constant-budget dictionaries of Appendix F.1 were not used for these ablations, so the SAE floors quoted in this paragraph are those of the original schedule and differ from the constant-budget floors of §5.2. A supervised logistic probe over the full 16k-dimensional SAE code reaches mean AUROC 0.856 at the final checkpoint, above the top-5 readout (0.841 on the seed average) and below the linear probe on activations (0.884). It flags the same two concepts at step 1 as the top-5 readout, but its refusal floor is 0.98 against 0.84 for the top-5 readout. The small number of flags is thus shared by a learned probe on the SAE code and is not specific to the top-5 readout, whereas the lower refusal floor is specific to the top-5 readout, which averages five features with equal weights. Concatenating activations and SAE code before probing reaches 0.891 and inherits the linear probe's nine flags, so on average the concatenation adds little for reading and does not reduce the flags. A per-concept readout selector that picks between the linear probe and the SAE top-5 readout by inner grouped cross-validation on the training folds reaches 0.887, within 0.004 of always choosing the probe, so adaptive selection is not worth its complexity at these gaps. None of these variants changes any conclusion of the paper, and we keep them out of the main comparison.

Layer 18 was audited with the original variable-budget dictionaries only, so its comparison with layer 12 uses those dictionaries at both layers; the constant-budget layer-12 dates reported in §5.3 are in Appendix F.1. At layer 18 the readout ordering is unchanged: the linear-probe and SAE top-5 means are lower (0.879 and 0.823–0.829 across seeds), the mean-difference mean is higher (0.671 against 0.614), and the probe has seven flags. Table 6 lists every concept's floor, corrected emergence step, and final AUROC at layer 18 in the format of Table 5, and Table 7 gives the group-median crossings of both layers under every readout and seed. Toxicity crosses at step 512 under all three dictionaries at both layers and at 1,000 under the probe, and bias crosses at 2,000 under every dictionary and never under the probe. Harm crosses at 1,000 under every dictionary at layer 12 but at 2,000 under the seed-123 dictionary at layer 18, and never under the probe at either layer; the epistemic group never crosses. The controls cross at 2,000 under the probe at layer 12 and at 4,000 at layer 18, and at 1,000–4,000 under the dictionaries depending on seed and layer; the refusal set crosses only under one dictionary at layer 12. The two layers therefore agree on the toxicity and bias dates and differ on harm for one seed and on the controls.

The OLMo-2-1B results reported here use layer 6, seed-42 dictionaries under the original schedule, and the 11 checkpoints of Appendix F. At step 0 the linear probe flags three concepts (politics, code, and refusal), a subset of the nine that it flags on Pythia-1.4B, and SAE top-5 flags refusal only. Table 8 gives the group steps at $\theta = 0.15$. Toxicity and harm cross at step 10,000 under both readouts. Bias crosses at 10,000 under SAE top-5 and at 80,000 under the probe, whereas on Pythia-1.4B it crosses under neither readout with the constant-budget dictionaries. The grid has no checkpoint between steps 300 and 10,000, so the order of toxicity and harm cannot be resolved under either readout, nor that of bias under SAE top-5, and under SAE top-5 the controls cross at step 300, before toxicity. OLMo-2-1B therefore supports the high probe floors of refusal and code and an early rise of toxicity, but not the order of the concept groups. At the final checkpoint the linear probe again reads best (mean AUROC 0.894 against 0.851 for SAE top-5, 0.815 for SAE top-1, and 0.843 for the mean-difference direction).

Table 9 adds uncertainty to the E1 comparison. The linear probe's advantage over the SAE top-5 readout is a paired difference of $+0.043$ AUROC on average over the 32 concepts (median $+0.039$), and a two-sided Wilcoxon signed-rank test rejects a zero median difference for every dictionary seed and for the seed average ($p < 0.001$). At the 0.02 margin the probe wins 19 or 20 concepts and loses 1 or 2 depending on the dictionary (20/10/2, 19/12/1, and 20/11/1 wins/ties/losses on seeds 42, 123, and 7). Because same-resource concepts share source texts, every aggregate is also given with a concept-level bootstrap interval and with a source-family cluster bootstrap interval that resamples six source families (BeaverTails, CrowS-Pairs, RealToxicityPrompts, and the control, refusal, and epistemic sets): the probe's mean AUROC 0.884 has cluster interval 0.810–0.956, and the best-or-

Table 6: **Per-concept floors, emergence steps, and final AUROC at layer 18** (Pythia-1.4B, grouped folds, original seed-42 dictionary; columns as in Table 5).

		Linear probe			SAE top-5		
Concept	Group	Floor	τ	Final	Floor	τ	Final
flirtation	toxicity	0.734	512	0.982	0.594	512	0.943
identity attack	toxicity	0.733	512	0.994	0.603	512	0.958
insult	toxicity	0.736	1,000	0.981	0.722	1,000	0.948
profanity	toxicity	0.777	2,000	0.961	0.704	2,000	0.930
severe toxicity	toxicity	0.710	512	0.984	0.621	512	0.839
sexually explicit	toxicity	0.742	1,000	0.983	0.654	1,000	0.976
threat	toxicity	0.760	1,000	0.991	0.623	512	0.960
animal abuse	harm	0.784	4,000	0.932	0.468	512	0.917
discrimination	harm	0.795	never	0.932	0.618	512	0.883
drugs weapons	harm	0.788	never	0.907	0.584	1,000	0.796
financial crime	harm	0.752	4,000	0.929	0.630	1,000	0.819
hate speech	harm	0.758	never	0.883	0.603	1,000	0.814
misinformation	harm	0.486	never	0.538	0.479	never	0.533
politics	harm	0.814	never	0.907	0.677	1,000	0.873
privacy	harm	0.808	never	0.944	0.692	4,000	0.874
self harm	harm	0.620	2,000	0.840	0.501	4,000	0.669
sexual	harm	0.698	2,000	0.949	0.589	1,000	0.906
terrorism	harm	0.742	never	0.869	0.568	4,000	0.763
unethical	harm	0.589	4,000	0.783	0.575	never	0.710
violence	harm	0.695	8,000	0.873	0.614	1,000	0.814
age	bias	0.618	never	0.699	0.602	never	0.575
gender	bias	0.652	never	0.655	0.568	2,000	0.615
nationality	bias	0.678	16,000	0.844	0.512	1,000	0.768
race	bias	0.840	never	0.890	0.568	2,000	0.791
religion	bias	0.806	never	0.871	0.532	2,000	0.856
socioeconomic	bias	0.786	never	0.851	0.573	8,000	0.757
refusal	refusal	0.998	never	1.000	0.889	never	0.993
sycophancy	epistemic	0.510	never	0.556	0.516	never	0.531
truthfulness	epistemic	0.629	never	0.717	0.544	1,000	0.726
code	control	1.000	never	1.000	1.000	never	1.000
science	control	0.802	512	1.000	0.639	512	0.998
sentiment	control	0.619	4,000	0.888	0.565	2,000	0.808

Table 7: **Group-median emergence steps at both layers under the original dictionaries.** The two-checkpoint crossing at $\theta = 0.15$ of each group's median floor-corrected trajectory under the linear probe and the three SAE top-5 dictionaries of the original variable-budget schedule (never = no sustained crossing); the constant-budget layer-12 values are in Table 14.

	Layer 12				Layer 18			
Group	Probe	SAE 42	SAE 123	SAE 7	Probe	SAE 42	SAE 123	SAE 7
toxicity	1,000	512	512	512	1,000	512	512	512
harm	never	1,000	1,000	1,000	never	1,000	2,000	1,000
bias	never	2,000	2,000	2,000	never	2,000	2,000	2,000
refusal	never	2,000	never	never	never	never	never	never
epistemic	never	never	never	never	never	never	never	never
control	2,000	2,000	1,000	2,000	4,000	2,000	4,000	2,000

tied count 31 has cluster interval 28.6–32 (count statistics are rescaled to 32 concepts). Table 10 treats the scale series the same way. The mean probe AUROC rises from 0.884 at 1.4B to 0.889, 0.893, and 0.896 at 2.8B, 6.9B, and 12B. Spearman correlations of per-concept probe AUROC with 1.4B are 0.990, 0.992, and 0.983 for 2.8B, 6.9B, and 12B. The per-concept differences have mean $+0.005$ to $+0.012$ with standard deviation 0.011–0.019, and between 5 and 10 concepts move by more than 0.02. The set of concepts above AUROC 0.7 is identical at 2.8B and 6.9B and differs at 12B by the gender axis alone, which crosses 0.7 only at 12B (0.64, 0.65, 0.68, and 0.70 from 1.4B to 12B).

The group steps and threshold details that §5.3 leaves out are as follows (every group step is in Table 14). After toxicity, the order depends on the dictionary: the controls cross at τ^{probe}=2,000 and at τ^{sae} of 1,000, 4,000, and 4,000. The epistemic group never crosses, a null result (final

Table 8: **Group-median emergence steps on OLMo-2-1B** (layer 6, $\theta = 0.15$, two-checkpoint rule; seed-42 dictionaries under the original schedule; never = no sustained crossing). The first three checkpoints of the grid are steps 0, 300, and 10,000, so steps of 10,000 are not ordered among themselves.

Group	Linear probe	SAE top-5
toxicity	10,000	10,000
harm	10,000	10,000
bias	80,000	10,000
controls	10,000	300
epistemic	never	80,000
refusal	never	10,000

Table 9: **Uncertainty for the readout comparison** (final Pythia-1.4B checkpoint, layer 12, 32 concepts). Top: paired difference of held-out AUROC, linear probe minus SAE top-5, per dictionary seed and on the seed average, with the two-sided Wilcoxon signed-rank p and the wins / ties / losses at the 0.02 margin. Bottom: point estimates with 95% intervals from a concept-level bootstrap and from a source-family cluster bootstrap (count statistics rescaled to 32 concepts under cluster resampling).

Seed	Mean diff.	Median diff.	Wilcoxon p	Wins / ties / losses
seed 42	+0.045	+0.043	<0.001	20 / 10 / 2
seed 123	+0.042	+0.031	<0.001	19 / 12 / 1
seed 7	+0.042	+0.031	<0.001	20 / 11 / 1
seed average	+0.043	+0.039	<0.001	19 / 12 / 1

Statistic	Point	Concept bootstrap	Cluster bootstrap
mean AUROC, linear probe	0.884	[0.839, 0.923]	[0.810, 0.956]
mean AUROC, SAE top-5 (seed average)	0.841	[0.798, 0.881]	[0.778, 0.932]
mean AUROC, mean-difference	0.614	[0.582, 0.653]	[0.568, 0.703]
probe wins (margin 0.02)	19.0	[14.0, 24.0]	[8.5, 24.2]
probe best or tied (margin 0.02)	31.0	[29.0, 32.0]	[28.6, 32.0]

medians $\Delta \leq 0.09$ under both readouts for every seed). For harm, at $\theta = 0.15$ the largest gain the probe median holds over two successive checkpoints is 0.139, just short of θ, and it crosses in 26% of resamples against 61–80% under the dictionaries; at $\theta = 0.10$ both cross (the dictionaries at 1,000, the probe at 4,000), and at $\theta = 0.20$ the dictionaries keep 11–17% against 0.2%. Bias (probe median 0.054, SAE medians 0.08–0.11) crosses under the SAE readout alone at $\theta = 0.10$ (2,000–8,000; 64–78% of resamples against 9%) and under neither at 0.15 (28–32% against 0.3%). The E3 crossings carry fold-level uncertainty estimated by a hierarchical bootstrap on the constant-budget trajectories: concepts are resampled within each group, every drawn concept contributes a trajectory rebuilt from its four stored fold AUROCs resampled with replacement, the floor is recomputed per draw, and the two-checkpoint rule of Definition 2 is applied to each of 2,000 resampled group medians. The toxicity crossing's 95% interval spans 512–4,000 across dictionaries (512–1,000 under the probe); the harm intervals run from 1,000 to never under every dictionary, with 20–39% of resamples never crossing, and the bias medians never cross at the point estimate and fail to cross in 68–72% of resamples. The probe medians for harm and bias fail to cross in 74% and 99.7% of resamples, so the readout-dependent dating of E3 is not a fold artefact. Table 11 extends this to every group and to $\theta \in \{0.10, 0.15, 0.20\}$ from the same resamples, and Figure 3(c) plots the harm and bias rows on a finer grid: a resample sustains a crossing at θ exactly when the largest gain its median holds over two successive checkpoints exceeds θ, so one pass over the draws gives the whole profile. On that grid the two readouts tie for harm below $\theta \approx 0.08$, where both cross at step 1,000, and the seed-123 dictionary falls behind the probe at $\theta = 0.115$ (same crossing step, fewer resamples); for bias they tie above 0.23, where neither crosses. At $\theta = 0.10$ the epistemic medians cross in 22–42% of resamples under the dictionaries and 30% under the probe although their point estimates do not, and the refusal set never crosses. The control interval is wide because that group holds three concepts and one is born at ceiling, leaving its median fragile under cluster resampling. Folds share training data and same-resource concepts share sources, so these intervals are read as descriptive stability summaries rather than exact-coverage inference. A released θ sweep replays both dating rules at $\theta \in \{0.10, 0.15, 0.20\}$ over all six groups and dictionaries: a probe median crossing before its SAE median occurs at three point estimates (the controls under two dictionaries at $\theta = 0.15$ and toxicity under one at $\theta = 0.20$), its bootstrap probability peaks at 0.42 in the toxicity

Table 10: **Scale series against Pythia-1.4B** (per-concept linear-probe AUROC at the final checkpoint, 32 concepts). Pearson and Spearman correlations with source-family cluster bootstrap 95% intervals; the per-concept difference (larger model minus 1.4B) as mean (SD) and as min / max; the number of concepts moving by more than 0.02; and the number of concepts on which the two models agree about exceeding AUROC 0.7.

Model	Mean	Pearson [CI]	Spearman [CI]	Δ mean (SD)	Δ min / max	$\lvert\Delta\rvert > 0.02$	Agree at 0.7
2.8B	0.889	0.996 [0.994, 0.999]	0.990 [0.922, 1.000]	+0.005 (0.011)	-0.011 / +0.037	5	32
6.9B	0.893	0.989 [0.983, 0.997]	0.992 [0.923, 1.000]	+0.009 (0.018)	-0.048 / +0.047	10	32
12B	0.896	0.989 [0.984, 0.997]	0.983 [0.871, 0.998]	+0.012 (0.019)	-0.009 / +0.064	8	31

setting whose two medians tie at step 1,000, and dropping the two-checkpoint requirement moves the seed-42 harm crossing at $\theta = 0.15$ from 32,000 to 8,000 and admits single-checkpoint epistemic excursions on two dictionaries at $\theta = 0.10$. Under the original variable-budget dictionaries the same bootstrap placed the harm and bias SAE crossings at 512–16,000 and 2,000–4,000 with at most 4% of resamples never crossing, the seed-level counterpart of the budget sensitivity in Appendix F.1. The same scripts recompute every table here.

Table 11: **Crossing probability by threshold** (constant-budget layer-12 trajectories, 2,000 hierarchical bootstrap resamples). Each cell gives the percentage of resamples in which the group median sustains a crossing at θ: linear probe, then the range over the three SAE top-5 dictionaries.

Group	$\theta = 0.10$	$\theta = 0.15$	$\theta = 0.20$
toxicity	100.0; 100.0–100.0	100.0; 100.0–100.0	100.0; 100.0–100.0
harm	100.0; 99.3–100.0	26.0; 61.3–79.9	0.2; 10.9–16.6
bias	8.6; 64.0–78.3	0.3; 27.9–31.7	0.0; 5.2–8.4
controls	75.0; 73.7–74.7	73.1; 73.7–74.7	32.3; 71.0–74.7
epistemic	29.5; 22.2–41.8	1.0; 0.4–1.3	0.0; 0.0–0.0
refusal	0.0; 0.0–0.0	0.0; 0.0–0.0	0.0; 0.0–0.0

E Probe Set Construction Details

This appendix describes how the 32 probe sets summarised in §4 are built, and reports the text-only classifiers that §5.2 cites. Table 12 lists the six concept groups with their sources and the contrast that defines each label.

Table 12: **The taxonomy of the 32 concepts.** Each concept is a balanced binary probe set with a frozen source-group prompt list; negatives are chosen to isolate the concept rather than a broader harmfulness signal.

Group	# concepts	Source	Positive vs. negative
Harm category	13	BeaverTails	category prompts vs. other-harm prompts
Toxicity axis	7	RealToxicityPrompts	high vs. low attribute score
Social bias	6	CrowS-Pairs	stereotypical vs. anti-stereotypical
Refusal	1	AdvBench / Dolly-15k	harmful requests vs. natural instructions
Epistemic	2	TruthfulQA / MWE	true vs. false; sycophantic vs. not
Control	3	SST-2 / AG News / CodeParrot	non-safety calibration

Every probe set is balanced, frozen to JSON before any extraction, and capped at 300 examples per side; we include a concept only if both sides retain at least 85 examples, the pre-registered minimum, which one further harm category (child abuse) fell below, so it is excluded. Harm-category positives are unique BeaverTails prompts carrying the category flag; negatives are sampled from prompts flagged *only* with other harm categories, so the contrast isolates category identity. Toxicity-axis pairs take the prompts scoring above 0.5 on the target attribute (the Perspective API scores released with RealToxicityPrompts) against prompts below 0.05 (the earlier toxicity probe set, used only as the example trajectory of Figure 6 and kept outside all aggregates, keeps its original 0.75/0.1 thresholds for comparability). CrowS-Pairs axes pair stereotypical with anti-stereotypical sentences per bias type; on the race axis the per-side cap leaves 183 complete pairs plus 234 singleton groups, so pairwise analyses apply to the complete-pair subset. Refusal pairs AdvBench harmful requests with length-filtered natural instructions from Dolly-15k that have no context field. TruthfulQA pairs question-plus-correct-answer with question-plus-incorrect-answer strings; sycophancy expands the embedded answer options of the model-written evaluation (MWE) items so the probe sees option text rather than option letters. Controls pair SST-2 positive with negative sentences, AG News science/technology with sports items, and CodeParrot Python files (first 600 characters) with AG

News world items. Residuals and tokenwise SAE features are each mean-pooled over up to 256 non-padding tokens. Concepts built from the same resource share source texts (BeaverTails categories reuse prompts across axes, as do RealToxicityPrompts attributes), so concepts are not statistically independent units; we therefore report per-concept quantities with per-concept uncertainty and avoid pooled cross-concept confidence claims.

Three text-only floor controls accompany the release, evaluated under the identical frozen fold contract with no model access: logistic regression on word-level TF-IDF (1–2 grams), on character 3–5 gram TF-IDF, and on length statistics alone. Word- and character-level TF-IDF exceed AUROC 0.8 on 23 and 24 of the 32 audited concepts respectively, including all nine probe-flagged ones, while the length-only control clears it solely on the code control. The words of the probe texts therefore allow more separation than the readouts reach at the first checkpoint. This is consistent with reading a flag as separability that the words already allow. A flag marks labels that a readout already separates at the first checkpoint, and we read a floor as what that readout extracts there, not as a complete account of lexical information.

F SAE Training Details and Ledger

This appendix lists the checkpoints, the SAE hyperparameters, and the compute ledger behind §4. Every SAE writes a wall-clock ledger entry; the 113 new SAEs consumed 23.2 GPU-hours and the 36 constant-budget dictionaries 13.3, so the marginal cost per concept trajectory is minutes. Appendix F.1 then reports the constant-budget retraining behind the layer-12 trajectory of §5.3. Dictionaries and readouts are computed at the 20 Pythia-1.4B checkpoints (steps 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1,000, 2,000, 4,000, 8,000, 16,000, 32,000, 64,000, 96,000, 128,000, and 143,000), at the 11 OLMo-2-1B stage-1 checkpoints (steps 0, 300, 10,000, 20,000, 40,000, 80,000, 150,000, 300,000, 610,000, 1,220,000, and 1,907,359), and at steps 8,000, 32,000, and 143,000 of Pythia-2.8B and 6.9B (layer 16) and 12B (layer 18). All SAEs are TopK autoencoders (Eq. 4) with k=32, dictionary expansion $8\times$ the residual width (16,384 for 1.4B-class models, 20,480 for 2.8B, 32,768 for 6.9B, 40,960 for 12B), decoder columns initialised at norm 0.1, and expected-average input normalisation, trained with Adam at learning rate 3×10^{-4} under a cosine schedule with 1,000 warm-up steps, batch size 2,048 tokens, and context length 512 on Pile text. Under the original Pythia schedule, late checkpoints receive 10^7 training tokens and early ones fewer (10^6 at steps ≤ 16, 2×10^6 at steps 32–128, and 3.33×10^6 at steps 256–1,000); Appendix F.1 removes this difference for the layer-12 Pythia-1.4B dictionaries. The OLMo-2 dictionaries receive 10^6 tokens at step 0, 3×10^6 at step 300, and 10^7 from step 10,000 on. Pythia-6.9B is loaded in bfloat16; activations are cast to float32 before encoding. Every run writes a ledger entry (seed, tokens, wall-clock, host); the totals below are incremental, since 40 reused dictionaries predate the ledger. The measured totals are 10.8 GPU-hours for the 80-SAE Pythia-1.4B multi-seed grid, 3.2 for the 22 OLMo-2 SAEs, and 3.6 for the six 2.8B/6.9B scale SAEs, on single RTX 4090 cards, plus 4.6 for the three bfloat16 Pythia-12B SAEs and 1.0 for the two 2.8B seed retrains on RTX A6000s. The 36 constant-budget dictionaries of Appendix F.1 add 13.3 hours of summed run wall-clock on RTX 4090 and A6000 cards; two dictionaries trained concurrently on each card, so this sum is not exclusive device time. The early-strength predictor of Appendix G uses t_{early}=8,000. The SAE, random, and INLP sensitivities use m=5 directions, the probe and mean-difference conditions use one, and LEACE is an affine eraser. The grouped protocol uses K=4 stratified folds throughout.

F.1 Constant-Budget Retraining of the Early Dictionaries

The original schedule trains an early dictionary on fewer tokens than a late one, so dictionary quality co-varies with the checkpoint and could shift the SAE floors and dates. The pre-registered control retrains every layer-12 Pythia-1.4B dictionary at steps 1 to 1,000 on 10^7 tokens for seeds 42, 123, and 7 (33 dictionaries) with every other hyperparameter unchanged and, as a reverse control, retrains step 2,000 on 3.33×10^6 tokens (3 dictionaries); Table 13 lists both schedules. The seed-42 retraining is a fresh run rather than the original seed-42 lineage. The constant-budget trajectory of each seed takes steps up to 1,000 from the retrained dictionaries and steps from 2,000 on from the original ones, which already have the 10^7 budget, and its floor is the retrained step-1 dictionary. Extraction reproduced the stored residuals and labels bit for bit (816 hash pairs), the dictionary-independent readouts reproduced the stored values at every shared step (1,224 comparisons), and an independent recomputation of the group crossings agreed with the analysis file. The pre-registered rule replaces the reported timeline with the constant-budget one when any group's date moves on two or more seeds; ten of the twelve group–seed dates at $\theta = 0.15$ moved (the pre-registered groups toxicity,

harm, bias, and controls on three seeds, Table 14; refusal and the epistemic group were outside the rule), so the constant-budget trajectory is the one reported in §5.3, Figure 3, and Table 5, and the original schedule is the sensitivity analysis of this appendix. Under the original schedule harm crossed at 1,000 and bias at 2,000 on every seed, and of the ten dates that moved, nine later or to never and one earlier.

Table 13: **Token budget per dictionary under the two schedules** (Pythia-1.4B, layer 12, every seed).

Checkpoint steps	Original schedule	Constant budget
1–16	10^6	10^7 (retrained)
32–128	2×10^6	10^7 (retrained)
256–1,000	3.33×10^6	10^7 (retrained)
2,000	10^7	10^7 (original); 3.33×10^6 reverse control
4,000–143,000	10^7	10^7 (original)

Table 14: **Group-median emergence steps under the original and the constant-budget dictionaries** (Pythia-1.4B, layer 12, SAE top-5, three thresholds). Each SAE cell reads original → constant budget; the probe column is dictionary-independent; never = no sustained crossing on the checkpoint grid, which does not mean that the concept is absent.

Threshold	Group	Probe	SAE seed 42	SAE seed 123	SAE seed 7
$\theta = 0.10$	toxicity	512	512 → 512	512 → 512	512 → 512
$\theta = 0.10$	harm	4,000	1,000 → 1,000	1,000 → 1,000	1,000 → 1,000
$\theta = 0.10$	bias	never	2,000 → 4,000	2,000 → 8,000	2,000 → 2,000
$\theta = 0.10$	controls	2,000	1,000 → 1,000	1,000 → 2,000	2,000 → 1,000
$\theta = 0.10$	epistemic	never	64,000 → never	4,000 → never	8,000 → never
$\theta = 0.10$	refusal	never	128 → never	128 → never	512 → never
$\theta = 0.15$	toxicity	1,000	512 → 1,000	512 → 512	512 → 512
$\theta = 0.15$	harm	never	1,000 → 32,000	1,000 → 4,000	1,000 → 2,000
$\theta = 0.15$	bias	never	2,000 → never	2,000 → never	2,000 → never
$\theta = 0.15$	controls	2,000	2,000 → 1,000	1,000 → 4,000	2,000 → 4,000
$\theta = 0.15$	epistemic	never	never → never	never → never	never → never
$\theta = 0.15$	refusal	never	2,000 → never	never → never	never → never
$\theta = 0.20$	toxicity	1,000	1,000 → 2,000	1,000 → 1,000	1,000 → 1,000
$\theta = 0.20$	harm	never	96,000 → never	4,000 → never	never → never
$\theta = 0.20$	bias	never	2,000 → never	2,000 → never	2,000 → never
$\theta = 0.20$	controls	never	2,000 → 2,000	4,000 → 4,000	4,000 → 4,000
$\theta = 0.20$	epistemic	never	never → never	never → never	never → never
$\theta = 0.20$	refusal	never	never → never	never → never	never → never

The direction of the change is uniform. The retrained early dictionaries read more at every step up to 1,000 (Table 15; the mean AUROC gain over the original dictionaries across the 32 audited concepts is $+0.01$ to $+0.09$ per step), and the smaller-budget step-2,000 dictionaries read less (-0.02 to -0.05), consistent with a budget effect on every seed. For seed 42 the comparison also changes the training lineage, so its four dates do not separate budget from lineage. From step 2,000 on the raw values are unchanged, so the higher step-1 floor shrinks the corrected gains: the mean SAE floor over the 32 concepts rises from 0.62 to 0.66–0.67, the refusal floor from 0.84–0.87 to 0.93–0.96, and the race, age, and socioeconomic floors from 0.53–0.58 to 0.66–0.77, which removes the bias crossings and delays the harm crossings, while the flag count stays at two (refusal and code) for every seed. Table 16 gives every audited concept's crossing under both schedules: of the 47 concept–seed crossings that moved, 41 moved later or to never and 6 moved earlier.

The dictionary-independent readouts are recomputed from the stored activations with the frozen folds; they reproduce the stored values exactly only under the BLAS thread count of the original run, which the released readout files record. All 36 runs write the same ledger entry as the other dictionaries.

Table 15: **Per-step effect of the token budget** (SAE top-5, 32 audited concepts). Mean (median) difference in held-out AUROC, retrained minus original dictionary at the same step; the last row is the reverse control, in which the retrained step-2,000 dictionary has the smaller budget.

Step	Seed 42	Seed 123	Seed 7	Comparison
1	+0.039 (+0.020)	+0.053 (+0.034)	+0.043 (+0.027)	10^7 vs. original
2	+0.069 (+0.065)	+0.062 (+0.047)	+0.043 (+0.029)	10^7 vs. original
4	+0.059 (+0.059)	+0.054 (+0.035)	+0.068 (+0.057)	10^7 vs. original
8	+0.073 (+0.077)	+0.086 (+0.088)	+0.065 (+0.062)	10^7 vs. original
16	+0.091 (+0.098)	+0.092 (+0.105)	+0.084 (+0.074)	10^7 vs. original
32	+0.051 (+0.042)	+0.036 (+0.025)	+0.044 (+0.029)	10^7 vs. original
64	+0.022 (+0.019)	+0.010 (+0.000)	+0.026 (+0.017)	10^7 vs. original
128	+0.009 (-0.001)	+0.025 (+0.008)	+0.012 (+0.003)	10^7 vs. original
256	+0.008 (-0.003)	+0.007 (-0.009)	+0.018 (+0.001)	10^7 vs. original
512	+0.021 (+0.005)	+0.020 (+0.004)	+0.012 (+0.008)	10^7 vs. original
1,000	+0.030 (+0.009)	+0.018 (-0.000)	+0.026 (+0.012)	10^7 vs. original
2,000	-0.048 (-0.029)	-0.024 (-0.021)	-0.032 (-0.029)	3.33×10^6 vs. 10^7

Table 16: **Per-concept emergence steps under the original and the constant-budget dictionaries** (SAE top-5, $\theta = 0.15$). Floor is the step-1 AUROC of the seed-42 dictionary, original $\rightarrow$ constant budget; each seed column reads original $\rightarrow$ constant budget.

Concept	Group	Floor (seed 42)	Seed 42	Seed 123	Seed 7
flirtation	toxicity	0.60 → 0.60	256 → 512	512 → 512	512 → 512
identity attack	toxicity	0.62 → 0.64	512 → 512	512 → 512	512 → 512
insult	toxicity	0.73 → 0.73	2,000 → 2,000	2,000 → 2,000	4,000 → 4,000
profanity	toxicity	0.71 → 0.72	2,000 → 2,000	4,000 → 4,000	4,000 → 4,000
severe toxicity	toxicity	0.65 → 0.66	512 → 512	512 → 512	512 → 512
sexually explicit	toxicity	0.66 → 0.68	512 → 2,000	1,000 → 1,000	1,000 → 2,000
threat	toxicity	0.61 → 0.63	512 → 1,000	512 → 512	512 → 512
animal abuse	harm	0.54 → 0.67	512 → 1,000	512 → 1,000	1,000 → 1,000
discrimination	harm	0.65 → 0.73	512 → 64,000	512 → 64,000	1,000 → 2,000
drugs weapons	harm	0.61 → 0.59	1,000 → 1,000	2,000 → 1,000	2,000 → 1,000
financial crime	harm	0.65 → 0.67	4,000 → 4,000	1,000 → 4,000	1,000 → never
hate speech	harm	0.61 → 0.62	1,000 → 1,000	4,000 → 4,000	512 → 1,000
misinformation	harm	0.53 → 0.48	never → never	never → never	never → never
politics	harm	0.65 → 0.72	512 → never	256 → 4,000	512 → never
privacy	harm	0.68 → 0.69	never → never	64,000 → never	4,000 → 4,000
self harm	harm	0.49 → 0.44	512 → 16	1,000 → 512	1,000 → 1,000
sexual	harm	0.58 → 0.60	1,000 → 1,000	1,000 → 2,000	1,000 → 2,000
terrorism	harm	0.57 → 0.59	8,000 → 32,000	2,000 → never	2,000 → never
unethical	harm	0.60 → 0.57	never → never	never → never	never → never
violence	harm	0.64 → 0.69	2,000 → never	1,000 → never	never → never
age	bias	0.54 → 0.72	8,000 → never	64,000 → never	64,000 → never
gender	bias	0.59 → 0.61	2,000 → 2,000	2,000 → never	2,000 → never
nationality	bias	0.53 → 0.67	2,000 → 8,000	2,000 → 2,000	2,000 → 2,000
race	bias	0.57 → 0.77	2,000 → never	2,000 → never	2,000 → never
religion	bias	0.57 → 0.57	2,000 → 1,000	2,000 → 2,000	2,000 → 2,000
socioeconomic	bias	0.53 → 0.73	2,000 → never	4,000 → never	2,000 → never
refusal	refusal	0.84 → 0.93	2,000 → never	never → never	never → never
sycophancy	epistemic	0.52 → 0.51	never → never	never → never	never → never
truthfulness	epistemic	0.58 → 0.63	96,000 → never	4,000 → never	128,000 → never
code	controls	1.00 → 1.00	never → never	never → never	never → never
science	controls	0.65 → 0.67	512 → 512	512 → 512	512 → 512
sentiment	controls	0.54 → 0.54	2,000 → 1,000	1,000 → 4,000	2,000 → 4,000

G Steering Baseline Details

This appendix supports §5.5. Appendix G.1 specifies the nine edits of Table 3 and the probe-level measurements of Table 17, and Appendix G.2 describes how settings are selected on validation prompts and certified on test prompts. The experiments that §5.5 summarises are the pre-registered confirmatory rounds on new prompts, insult, OLMo-2-1B, and threat (Appendices G.3, G.6, and G.7, Table 19) and the control with subspaces selected for unrelated concepts (Appendix G.4). Three parts support the Discussion: the sensitivity to dictionary seeds, thresholds,

and sampling (Appendix G.8), held-out perplexity (Appendix G.10), and what the selected features read (Appendix G.11). Two analyses appear only here: a semantic rescoring of the completions (Appendix G.5) and the early-strength predictor (Appendix G.9).

G.1 Protocol and Baselines

All behavioural comparisons run in one regime: unbatched greedy decoding of 50 new tokens, scored by Detoxify (unbiased), with every steering direction, subspace, or eraser fit on the stored probe activations only. Unlike the fold-wise subspaces of Definition 3, the generation-time subspace is fit once on all stored probe activations; the steering prompts are external to every probe set, so no fold separation is needed there. Prompts come from a frozen RealToxicityPrompts prompt list disjoint from every probe set and from the 200-prompt development runs; 200 validation prompts per class select each method's setting (its rank, strength, or head count), which we call an operating point, and 500 test prompts per class are generated exactly once (§5.5). Subspace removal applies the projection of Eq. 8 at each token's post-block residual during prompt processing and decoding (block 12 at 1.4B, block 16 at 2.8B); its partial-strength variant interpolates,

$$\phi' \;=\; \phi - \lambda\, QQ^{\top}\phi, \qquad \lambda \in [0,1], \tag{16}$$

recovering full removal at λ=1. The activation-addition baseline in the style of ActAdd and contrastive activation addition (Turner et al., 2023; Rimsky et al., 2024) subtracts the mean-difference vector at full or half strength,

$$\phi' \;=\; \phi - \beta\, w_c, \qquad \beta \in \{1, \tfrac{1}{2}\}, \tag{17}$$

a translation rather than a projection, applied at every position, where w_c is the positive-minus-negative mean of the stored probe activations. The LEACE baseline applies the closed-form affine eraser of Belrose et al. (2023) for a binary concept,

$$\phi' \;=\; \phi - \Sigma^{1/2} u u^{\top} W(\phi - \mu), \qquad u = \frac{W w_c}{\|W w_c\|}, \quad W = (\Sigma^{1/2})^{+}, \tag{18}$$

with μ, Σ the mean and covariance of the stored probe activations, Moore–Penrose roots taken on the data subspace (eigenvalues below $10^{-10}\lambda_{\max}$ truncated), and u the unit whitened class-mean direction; the implementation is checked in-run by verifying that a logistic probe on the erased stored activations scores near chance. Steering maps fit on pooled probe representations act on each token's residual. The supervised comparison was specified as five iteratively refit mean-difference directions, each computed after projecting out its predecessors. Projecting out the class-mean difference makes the two class means coincide, so the refit terminates after the first direction (the second iterate has norm below 10^{-9} on every activation cache), and the supervised comparison is the rank-1 mean-difference projection in every run reported here; the SAE and random subspaces have full column rank. The RepE reading vector is the first principal component of differences between randomly paired positive and negative cache activations, applied as a translation whose length is the gap between the two class means along that component. SAE feature clamping subtracts, at every token, each selected decoder direction scaled by that feature's encoder activation, so it removes the features' contribution only where they fire. The second, independent toxicity classifier used for rescoring is the RoBERTa toxicity classifier released by s-nlp.

The representation-level sensitivities of §5.5 add three conditions to the SAE, mean-difference, and random subspaces (Table 17): the clean probe's own weight direction mapped back to activation space (rank 1), five iterative nullspace-projection directions fit as in the INLP baseline (a logistic probe is fit on the labels, its direction is projected out, and the fit is repeated five times), and the LEACE eraser applied as an affine map. Each is fit on the training fold and applied to the held-out fold (fixed) or to the training fold before refitting (fresh), exactly as Definition 3. Removing the probe's own direction makes the fixed probe's score constant, so its AUROC is exactly 0.5 by the tie convention of Eq. 1; the projection is carried out in double precision, because single-precision rounding leaves a spurious ranking of the constant scores. Averaged over the 32 audited concepts, for the rank-5 SAE subspace the fixed probe loses 0.15 while a refit probe loses only 0.009; after removing the probe's own direction a refit probe recovers to 0.82, and the five directions of iterative nullspace projection (INLP) still leave a refit at 0.79, whereas removing the mean-difference direction or applying LEACE holds the refit at chance (0.50).

G.2 Initial Toxicity Round: Selection and Certification

In this appendix k^* denotes the validation-selected value of the removal rank m of Definition 3, distinct from the TopK sparsity k of Eq. 4; α is the ITI addition strength and K its head count, distinct from the tail probability of Proposition 3 and the fold count of Definition 1. On the validation

Table 17: **Representation-level sensitivities of six subspaces** (Pythia-1.4B, layer 12, final checkpoint, 32 audited concepts, grouped folds). σ_{fix} and σ_{fresh} are the held-out AUROC lost by the original and by a refit probe (Eq. 9), averaged over folds and concepts; the SAE, mean-difference, and random rows average the three dictionary seeds' runs; the last column is the refit probe's AUROC against a clean mean of 0.884.

Removed subspace	Rank	σ_{fix}	σ_{fresh}	Refit AUROC
SAE top-5 subspace	5	0.153	0.009	0.875
Mean-difference direction	1	0.391	0.384	0.500
Random subspace	5	0.001	0.000	0.884
Probe's own direction	1	0.384	0.063	0.821
INLP directions	5	0.416	0.096	0.787
LEACE eraser	1	0.293	0.382	0.502

split the rank and the strength are each swept in one dimension around the full-strength rank-5 setting: the rank sweep runs at full strength, the strength sweep runs at rank 5, and no condition that changes both is generated. On Pythia-1.4B, rank-m SAE removal moves -17, -20, and -7% at $m = 1, 2, 3$, reaches -36% at m=5, and deepens to -42% at m=10 and -98% at m=20, where distinct-2 (the distinct-bigram fraction d_2 of Table 3: distinct word bigrams divided by all word bigrams, pooled over the completions of a condition) falls to 0.56 of the clean run's value. The diversity criterion of §5.5, distinct-2 at or above $0.85\times$ clean, excludes $m \geq 10$; validation fixes k^*=5. The partial strengths ($+1$, -2, -14% at $\lambda = 0.25, 0.5, 0.75$) all pass the in-run selection criteria, and λ=0.75 is selected as the deepest of them, full strength remaining the main setting. On Pythia-2.8B the same rule selects k^*=3, which the test split confirms at -47%. On the 1.4B test split the seed-42 subspace passes all three criteria at -38%, while of the other two seeds one (-35%) misses the benign criterion by confidence width and one (-29%) the diversity criterion by 0.01.

The released analysis script records how the SAE rank and strength of this initial round were selected. One message string inside the generation script still describes the distinct-2 threshold as applying to validation only. The criteria that apply are the ones stated in this appendix and in the pre-registration documents, and the released check enforces them. These SAE settings were selected on validation with a signed-rank (median) form of the benign criterion, computed inside the generation script, whereas the ITI settings were selected with the mean-based criterion from the start. Before any test file was read, we re-specified the benign criterion on the mean, which is the quantity that rises under heavy tails (one-sided 95% paired-bootstrap upper bound, 10,000 resamples). For the initial SAE selections this stricter rule is applied to the test results only. At n=200 the bootstrap upper bounds of the deepest validation settings cross the margin while their point estimates stay negative, so the criterion is judged on the n=500 test split, where the selected operating points pass it. The Holm correction of the significance criterion runs over the conditions of each test file other than the clean run and the random subspaces (14 conditions at 1.4B, 12 at 2.8B, and two in each ITI test file). The benign criterion is a per-condition bound without simultaneous coverage, so it screens individual rows and supports no family-wise claim. One benign prompt string recurs across the validation and test sides (699 of 700 benign entries unique); the prompt list is kept as frozen.

The ITI grid moves monotonically in both the head count and the strength (-6 to -51% over $K \in \{24, 48\}$, $\alpha \in \{5, 15\}$ at 1.4B), selecting K^*=48, α^*=15 at both sizes; its five-head setting (K=5), added post hoc to match the rank-5 budget and labelled exploratory, shows no detected effect on the test split. The supervised direction's benign leak (its failure of the benign criterion) persists at every operating point of the validation and test splits, while every SAE test point above half strength leaves benign toxicity at or below clean (the λ=0.5 sweep point sits marginally above and misses the certificate).

The RepE vector and the half-strength supervised direction at 2.8B increase toxicity (Table 3). The second scorer reproduces the method ordering of Table 3 (Spearman ρ=0.98 at 1.4B) and also rescored every benign completion, replaying the frozen margin rule on its own scale. All selected operating points certify (both sizes' SAE primaries and all confirmatory ITI points, with negative point estimates throughout), and the supervised direction fails at every point of both sizes. Two results depend on the scorer: the 1.4B seed-123 and seed-7 upper bounds cross the margin by confidence width on negative point estimates, and 2.8B LEACE, which fails the pre-registered criterion, clears it under the second scorer while 1.4B LEACE fails under both; the pre-registered Detoxify criterion remains the certificate of record.

G.3 Confirmatory Rounds: Fresh Prompts and Insult

A pre-registered confirmatory round, frozen and committed before any generation, replays the whole design under one rule: selection and certification are both the bootstrap mean margin with the distinct-2 criterion, recomputed from the existing validation outputs, and evaluated exactly once on fresh prompt sets not previously used, each verified disjoint from every earlier prompt list and from the stored probe activations. Table 19 reports the outcome. On fresh toxicity prompts the ordering of methods replicates at both sizes: the 2.8B rule-selected subspace (rank 3) carries the full certificate at -49%, ITI does at -43 and -49%, the seed-123 upper bound again misses by confidence width on a negative point estimate, the supervised direction again fails the benign and diversity criteria, and the ten random subspaces again stay within single digits. The rule's own 1.4B selection (the seed-7 subspace) is the one exception: it cuts -33% with benign intact but its test distinct-2 lands at $0.82\times$ clean, under the 0.85 criterion its own rule sets, while the pre-registered rank-5 subspace passes all three criteria at -30%; the fresh round thereby catches a selection made on confidence-limited validation data, which is what the criterion is for. On the second behaviour, the same protocol applied to the insult attribute selects rank 3 at both sizes and fully certifies -37 and -52%, ITI certifies -84 and -70% on the pre-registered criteria (the single selection rule rejects the 48-head point at 1.4B for its diversity cost and selects 24 heads, and the semantic rescoring below measures what these deepest cuts cost), the supervised direction fails the benign criterion in every setting, and LEACE certifies on insult at both sizes, at depths it never reaches on toxicity, with activation addition certified at 2.8B and rejected by the diversity criterion at 1.4B. The second scorer rescored the fresh toxicity round on both sides: every selected SAE and ITI point passes the benign margin in all eight combinations of scorer, prompt set, and size, while LEACE's benign certificate flips with the draw and the scorer (it passes the pre-registered criterion at 1.4B on fresh prompts and fails under the second scorer, and fails both at 2.8B). The insult round was rescored in the same way. There the second scorer's per-condition means track the primary at Spearman ρ=0.99 (1.4B) and 0.98 (2.8B), and every certified SAE and ITI point keeps one-sided significance on the steered side and the benign margin on the other. The supervised direction again removes nearly all scored insult (-96 and -99%, with the 1.4B benign side failing), and the 1.4B seed-123 subspace, whose upper bound already crossed by confidence width on toxicity, fails the margin here as well.

G.4 Specificity Control with Unrelated Concepts

The random-subspace comparison of §5.5 calibrates the projection for random directions, which remove an expected m/d of any fixed direction (Proposition 3). Dictionary directions are not random: on the stored toxicity activations, a rank-5 subspace of the seed-42 dictionary removes 3.8% of the mean-centred variance at 1.4B and 2.4% at 2.8B, against 0.26% and 0.21% on average for a Gaussian rank-5 subspace (about m/d). Across the toxicity and control subspaces, five active dictionary directions remove 2–8% of the activation variance (random rank-5: 0.2%). One exploratory control, run after the pre-registered rounds with its dictionary, rank, selection rule, and feature indices frozen and written to the repository before any generation, therefore asks whether the reduction of Table 19 follows from the toxicity features or from removing any five active dictionary directions. Four rank-5 subspaces of the same dictionary are removed at each size under the protocol of §5.5, on the fresh toxicity holdout of Table 19, against a clean run regenerated in the same session (its toxic-prompt mean matches the recorded one to four decimals). Three apply the toxicity subspace's own rule, features ranked by $|\mathrm{AUROC} - \frac{1}{2}|$ on the stored activations, to probe sets for other concepts, a sentiment set, a science-topic control set, and the gender pairs; a feature shared with the toxicity top five is replaced by the next-ranked one (one replacement for the sentiment set at 1.4B, two at 2.8B). The fourth pairs each toxicity feature with the feature of nearest activation frequency (the fraction of probe texts on which the pooled feature is positive) among those whose toxicity $|\mathrm{AUROC} - \frac{1}{2}|$ lies in the middle fifth of the dictionary (at 2.8B one toxicity feature fires on every probe text and has no close match). The science and frequency-matched subspaces carry no toxicity signal (their features rank 195th or lower for toxicity), the gender subspace little (features ranked 44th and 77th at 1.4B and 16th at 2.8B), and the sentiment subspace some by construction (features ranked sixth and eighth at 1.4B, twelfth at 2.8B); all lie at principal angles of 71–90 degrees to the toxicity subspace and remove 2.2–7.8% (1.4B) and 1.7–4.0% (2.8B) of the centred variance. Under Definition 3 on the toxicity probe set, none of the four removes toxicity information that a refit probe detects: their fresh-probe drops are no larger than 0.002 in magnitude at either size, against 0.026 and 0.036 for the toxicity subspace, and their fixed-probe drops are at most 0.012 and 0.005, against 0.23 and 0.15.

Table 18 reports generation. At 1.4B every control reduces toxic-prompt toxicity, by 21% (science) to 39% (frequency-matched), against 30% for the toxicity subspace on the same prompts; the sentiment, science, and gender subspaces pass all three criteria, and the frequency-matched subspace, which removes the most variance, fails the benign and diversity criteria. At 2.8B the sentiment, gender, and frequency-matched subspaces cut 18–19% and pass all three criteria, the science subspace changes toxicity by 2%, short of the significance criterion, and the toxicity subspace cuts 51%. Benign-prompt toxicity falls with every passing control (from 0.009 to 0.003 at 1.4B and from 0.007 to 0.002–0.004 at 2.8B), as it does under the toxicity subspace. Removing five active dictionary directions with little or no toxicity signal therefore moves the classifier by an amount the random null does not anticipate: at 1.4B the science subspace, whose features rank 221st or lower for toxicity, cuts 21% and the frequency-matched subspace 39%, as much as or more than the toxicity subspace; at 2.8B the controls reach about a third of its reduction. The certificate of §5.5 measures the size and the side effects of an edit and does not identify which features carry the behaviour; subspaces selected for unrelated concepts from the same dictionary are the null against which specificity can be judged, and by it the toxicity subspace's reduction exceeds unrelated subspaces only at 2.8B. Language quality moves with the edit as well: held-out perplexity (the protocol of Appendix G.10) rises by 20–39% under the three passing controls at 1.4B, against 26–38% under the toxicity subspace, and by 73–100% under the frequency-matched one; at 2.8B the controls cost 13–59% against 15–20% for the toxicity subspace while reducing toxicity less, so there the toxicity subspace is the deeper edit per unit of perplexity. The second toxicity classifier gives the same picture (shifts of -23 to -42% at 1.4B and -2 to -20% at 2.8B).

Table 18: **Rank-5 subspaces selected for unrelated concepts, removed on the fresh toxicity holdout.** Shift: relative change of mean toxic-prompt toxicity against a clean run regenerated in the same session; marks give the H / B / d_2 criteria of §5.5 in order, with H Holm-corrected over the four controls. The reference rows transcribe the fresh-holdout round of Table 19: its toxicity rank-5 subspace of the same dictionary and the range of its ten Gaussian random subspaces (outside the Holm family). The control was frozen before generation and is not pre-registered.

	Pythia-1.4B		Pythia-2.8B	
Removed subspace	Shift (%)	H/B/d_2	Shift (%)	H/B/d_2
Sentiment-5	-31	✓✓✓	-19	✓✓✓
Science-5	-21	✓✓✓	-2	✗✓✓
Gender-5	-30	✓✓✓	-18	✓✓✓
Frequency-matched-5	-39	✓✗✗	-19	✓✓✓
SAE-5 (toxicity, seed 42)	-30	✓✓✓	-51	✓✓✓
Random-5 (10 draws)	-4 to $+4$	–	-1 to $+5$	–

G.5 Semantic Rescoring

Because distinct-2 and length are lexical, every stored test completion was also rescored post hoc with a MiniLM sentence encoder (all-MiniLM-L6-v2, mean-pooled, released script), yielding two cosine similarities per prompt: *retention*, against the clean run's continuation of the same prompt, and *relevance*, against the prompt itself, reported relative to the clean run's own relevance. The rule-selected and pre-registered SAE points hold 86–96% of clean relevance on both behaviours (89–96% over the seven fully certified ones, and 86% for the rule's own 1.4B selection, which missed the diversity criterion) and toxicity ITI holds 89–90%, while the insult ITI points keep 64% (24 heads, 1.4B) and 79% (48 heads, 2.8B): the deepest cuts drift off topic in a way that the lexical criteria do not detect, which we flag as a caveat on the insult ITI depths. Relevance also varies with the ITI strength. On the rescored α=15 test points, 2.8B's 24 heads keep -60% at 86% relevance while neither 1.4B point holds relevance at $0.85\times$ clean (64% and 45%); rescoring the validation grid adds the α=5 half, where 1.4B's 24 heads cut -43% at 89% relevance. Deep (at least 40% reduction), semantically clean (relevance at or above $0.85\times$ clean) ITI points therefore exist at both sizes, and the α=15 selections traded relevance for depth because the frozen rule screens only lexical axes: a semantic criterion belongs in the selection rule, which we flag as protocol guidance. The collapsed supervised direction annihilates both similarities (retention 0.08–0.11), and random subspaces keep the clean run's relevance while their text drifts (retention 0.71–0.86), so the semantic axis separates the same families as the certificates. Human quality judgements remain future work; this axis is an automated proxy.

G.6 Second Model Family

A third pre-registered round applies the unchanged protocol to a second model family, OLMo-2-1B's final stage-1 checkpoint, with a fresh prompt list, the OLMo-2 dictionary and probe activations already used for the readout comparison, and a startup check that the generation-time hook reproduces the stored reading activations (cosine 1.0000 on eight probe texts). The protocol transfers: validation selects rank k^*=20 and the test split fully certifies it at -53% with distinct-2 above clean and 87% relevance, so an SAE subspace again gives the deepest reduction that passes the benign and diversity criteria. The model-specific structure does not: rank 5 cuts only -21% (certified but shallow; the deepest certified reduction needs rank 20 here, against ranks three to ten on Pythia), activation addition strengthens to -26%, and the supervised direction's primary-scorer leak does not appear. The second scorer brings back the pattern of the Pythia rounds: only the two SAE points keep their benign certificates under it, activation addition and the supervised direction lose theirs, and LEACE fails under both scorers on OLMo-2, which adds to its record on Pythia of depending on the prompt draw and the scorer. Head probes were never extracted for OLMo, so this round carries no ITI experiment.

G.7 Third Behaviour: Threat

A fourth pre-registered round applies the protocol to a third behaviour, threat, chosen as the attribute farthest from toxicity among those the primary scorer covers (prompt-side rank correlation 0.62 over the scored RealToxicityPrompts pool, against 0.97 for insult; the one farther attribute, flirtation at 0.50, has no Detoxify head), with a fresh list of 1,400 prompts verified disjoint from every earlier prompt list and the stored probe activations, the same scorer's threat head as primary, and the single selection rule unchanged. Validation selects rank k^*=3 at 1.4B and k^*=10 at 2.8B, a third pattern of selected ranks after 3–5 for toxicity and 20 for OLMo-2, and the test split certifies both at -47 and -82%, the deepest certified SAE selection of any round. All three 1.4B rank-5 dictionaries carry the full certificate (-63/-85/-59 for the original, seed-123, and seed-7 dictionaries), so all three 1.4B subspaces certify on threat (-59 to -85%), as does the pre-registered seed-42 subspace at 2.8B (-65%). The ITI strength is behaviour-dependent and the selection rule handles it: at 1.4B both α=15 points fail the validation diversity criterion, the rule selects 48 heads at α=5, and the test certifies -48% at distinct-2 above clean while the validation-passing runner-up (24 heads) shows no detected test effect; at 2.8B the rule selects 24 heads at α=15, the test certifies -48%, and the 48-head point is rejected by the diversity criterion. The certified set itself is behaviour-dependent: activation addition (-33/-39%) and LEACE (-44% at both sizes) carry full certificates on threat at depths they never reach on toxicity, INLP cuts -67 and -33% only below the diversity criterion, and the supervised direction keeps its benign leak at 1.4B (-87%) and inverts at 2.8B, amplifying threat $+595\%$ at full strength and $+852\%$ at half with benign and diversity failures. The ten random subspaces stay inside -11 to $+6\%$ (1.4B) and -2 to $+15\%$ (2.8B), three draws reach nominal $p<.05$ (the random conditions sit outside the pre-registered Holm family), and every subspace-removal run passes a check, made before generation on eight probe texts, that the hooked residual agrees in direction with the stored activation (minimum hook cosine 1.0000, saved with each output file).

Two post-hoc rescorings of the frozen threat completions follow, one with the independent classifier and one with the semantic measure. Under the independent classifier the 2.8B results replicate in full: per-condition means track the primary at Spearman 0.89, every certified point keeps one-sided significance with the replayed benign margin intact, and the supervised inversion reproduces at $+286$/$+375\%$ with a benign upper bound two orders of magnitude past the margin. At 1.4B the transfer is scorer-sensitive and we report it as measured: the selected ITI point and activation addition re-certify (-43 and -26%), the SAE subspaces and LEACE keep one-sided significance (-12 to -46%) while their benign upper bounds cross the replayed margin ($+0.0033$–$+0.0071$ against $+0.003$), and per-condition agreement drops to Spearman 0.69, the gap between a threat head and a general-toxicity scorer at the smaller size; the pre-registered threat-head criterion remains the certificate of record. The semantic rescoring shows the same costs for threat as for insult: certified moderate points hold 85–96% of clean prompt-relevance (ITI 93–94%), the two deepest SAE cuts pay for depth (the seed-123 subspace's -85% at 75%, the rank-10 selection's -82% at 71%), the 1.4B supervised collapse annihilates relevance (20%), and the 2.8B inversion amplifies on topic at half strength (101%) while full strength drifts (59%).

G.8 Seed Retraining, Threshold Sensitivity, and Sampling

A post-hoc seed-robustness experiment, disclosed as such (it reuses the toxicity fresh-holdout test split and performs no selection), retrains the 2.8B dictionary from seeds 123 and 7 under unchanged hyperparameters and evaluates the three fixed rank-5 subspaces once, on the machine that generated the frozen artifact. In short, a post-hoc 2.8B experiment with two retrained dictionaries certifies seed 123 at -39%, while seed 7's -97% cut collapses distinct-2 to $0.40\times$ clean and fails the diversity criterion, as detailed next. Determinism holds to the token: the re-run clean and seed-42 completions match the frozen test rows on all 2,000 strings, the re-extracted seed-42 features are bit-identical to the frozen cache, and the seed-42 subspace reproduces its frozen certificate to the digit. The seed-123 subspace carries the full certificate at -39%. The seed-7 subspace cuts -97% with the benign margin intact but collapses distinct-2 to $0.40\times$ clean (448 of 500 completions unique), and the criterion rejects it, so the loss of diversity was caught by the pre-registered criteria. Seed spread therefore differs between these two settings, two of three dictionaries certified at 2.8B against three of three on 1.4B threat. The rescores extend here: the second scorer reproduces the three subspaces' relative changes to within a point (-52/-39/-98%), and the semantic axis separates them the same way, both certified dictionaries at 91% of clean relevance and the criterion-rejected seed-7 subspace at 31%.

Whether a certificate replicates separates the method families: ITI's full certificate holds in every selected confirmatory setting, the rank-5 SAE subspace's in all six behaviour–size combinations (-30 to -65%), and LEACE's in no consistent pattern, which is what we set out to measure. Most headline certificates do not hinge on the exact thresholds. Re-reading the frozen artifacts over benign margins 0.001–0.010 and distinct-2 factors 0.80–0.90 leaves seventeen of the twenty certified headline settings (rule-selected, pre-registered rank-5, ITI, both OLMo SAE points, and the certified 2.8B retrain) unchanged at every grid point. Every headline setting that loses its certificate does so at the strictest diversity factor of 0.90 (the 2.8B toxicity rank-5, threat rule-selected, and seed-123 retrain subspaces, distinct-2 ratios 0.85–0.88). No headline setting changes when only the benign margin varies, and the only SAE or ITI setting that gains a certificate is the disclosed toxicity seed-7 point at the factor 0.80 (outside the headline set, the two LEACE settings enter at the loosest margin); the collapsed retrain enters nowhere on the grid. Under nucleus sampling (top-p 0.9, 300 prompts per class) the selected SAE subspaces reach -64 to -85% across sizes and decoding seeds with their diversity cost vanishing, while the LEACE and supervised benign leaks surface in the point estimates.

Table 19: **Confirmatory rounds on the held-out behaviour grid** (single selection rule, pre-registered). Cells give the target-side shift in % and the three criteria in order Holm / benign / distinct-2 ($0.85\times$ clean); ✓✓✓ is the full certificate. Rule selections: k^* = seed-7 rank 5 (toxicity 1.4B), rank 10 (threat 2.8B), rank 3 elsewhere; K^* = 48 heads except 24 at insult 1.4B and threat 2.8B, with ITI strength α=5 at threat 1.4B and α=15 elsewhere. Ten random subspaces per setting sit outside the pre-registered Holm family (nominal $p<.05$: one insult-1.4B and three threat draws). The rank-5 row is the pre-specified pre-registered secondary (the original dictionary at 1.4B, seed 42 at 2.8B); its test d_2 ratios are .90/.88/.92/.92/.95/.98 and ITI's are .91/.98/.97/.99/1.05/.96.

	Toxicity (fresh holdout)				Insult (second behaviour)				Threat (third behaviour)			
	1.4B		2.8B		1.4B		2.8B		1.4B		2.8B	
Method	Shift	H/B/d_2	Shift	H/B/d_2	Shift	H/B/d_2	Shift	H/B/d_2	Shift	H/B/d_2	Shift	H/B/d_2
SAE rule-selected (k^*)	-33	✓✓✗	-49	✓✓✓	-37	✓✓✓	-52	✓✓✓	-47	✓✓✓	-82	✓✓✓
SAE rank-5 subspace	-30	✓✓✓	-51	✓✓✓	-38	✓✓✓	-55	✓✓✓	-63	✓✓✓	-65	✓✓✓
ITI (K^* heads)	-43	✓✓✓	-49	✓✓✓	-84	✓✓✓	-70	✓✓✓	-48	✓✓✓	-48	✓✓✓
LEACE	-23	✓✓✓	-20	✓✗✓	-41	✓✓✓	-46	✓✓✓	-44	✓✓✓	-44	✓✓✓
ActAdd (full)	-8	✗✓✓	-15	✓✓✓	-10	✓✓✗	-29	✓✓✓	-33	✓✓✓	-39	✓✓✓
Superv. (full)	-87	✓✗✗	$+28$	✗✗✗	-97	✓✗✗	-97	✓✗✗	-87	✓✗✗	$+595$	✗✗✗

G.9 Early Strength and Final Sensitivity

Early strength is the floor-corrected gain Δ_c^R of the SAE top-5 readout at step 8,000, computed here with the original-schedule dictionaries and their floors. It is measured separately from the final-checkpoint intervention outcome and predicts the final subspace sensitivities of §5.5 at Spearman $\rho = 0.56$–0.59 for σ_{fix} and 0.49–0.56 for σ_{fresh} across seeds (n=32; concept-level and source-family cluster bootstrap CIs exclude zero, recomputed by a released script from the audited-concept artifact), replacing the earlier version's inflated four-concept r=0.64 that arose from selection-evaluation reuse.

G.10 Language Quality After Removal

The interventions of Table 3 (for ITI the 24- and 48-head settings, and three of the ten random subspaces) were re-applied at every token position of two fixed held-out texts, the first 300 documents of the Pile validation split and of the WikiText-103 test split truncated to 512 tokens (109,358 and 49,394 scored tokens), and scored by teacher-forced token negative log-likelihood; Table 20 reports perplexity and its ratio to the clean run for Pythia-1.4B (block 12) and 2.8B (block 16) at the final checkpoint. No threshold was pre-registered for this quantity, so it is reported as a cost rather than added to the criteria after the fact. The selected SAE subspaces of the initial round raise perplexity by 26–38% at 1.4B (k^*=5) and 11–14% at 2.8B (k^*=3), the three 1.4B dictionaries lying within 26–44%; partial strength roughly halves the cost (λ=0.75: 9–20%), feature clamping and random subspaces cost at most 3%, ITI costs 8–41% depending on the head count, LEACE costs 1–30%, INLP multiplies perplexity by 1.2–6.0, and the supervised direction by 11–29 at full strength. Perplexity measures next-token prediction on unedited text only; it does not certify generation quality, for which the benign and diversity criteria and the relevance measure above remain the evidence, and no human judgement was collected.

Table 20: **Held-out perplexity after each intervention** (final checkpoint; 300 documents each of the Pile validation and WikiText-103 test splits, 512 tokens per document). Cells give perplexity with the ratio to the clean run in parentheses; SAE rows name the dictionary seed, k^* the validation-selected rank, and λ the removal strength of Eq. 16; Superv. (half) removes the mean-difference direction at λ=0.5; the λ=0.5 SAE row is a supplementary operating point of the initial round.

	Pythia-1.4B		Pythia-2.8B	
Intervention	Pile	WikiText-103	Pile	WikiText-103
Clean (unmodified model)	7.91	22.51	7.18	19.33
SAE-5, seed 42	9.99 (1.26)	31.12 (1.38)	8.30 (1.15)	23.16 (1.20)
SAE-5, seed 123	10.34 (1.31)	32.35 (1.44)	–	–
SAE-5, seed 7	10.11 (1.28)	30.56 (1.36)	–	–
SAE, rule-selected rank (k^*=5 / 3)	9.99 (1.26)	31.12 (1.38)	8.00 (1.11)	21.98 (1.14)
SAE-5 at strength λ=0.75	9.02 (1.14)	27.08 (1.20)	7.83 (1.09)	21.47 (1.11)
SAE-5 at strength λ=0.5	8.44 (1.07)	24.75 (1.10)	7.51 (1.05)	20.31 (1.05)
SAE clamp	7.91 (1.00)	22.52 (1.00)	7.19 (1.00)	19.34 (1.00)
RepE vector	8.21 (1.04)	23.71 (1.05)	7.21 (1.00)	19.49 (1.01)
ActAdd (full)	8.26 (1.05)	23.94 (1.06)	7.26 (1.01)	19.64 (1.02)
ActAdd (half)	7.99 (1.01)	22.85 (1.02)	7.20 (1.00)	19.41 (1.00)
LEACE	10.29 (1.30)	26.82 (1.19)	7.23 (1.01)	19.52 (1.01)
INLP-5	26.10 (3.30)	135.32 (6.01)	8.73 (1.22)	25.95 (1.34)
Superv. (full)	190.88 (24.14)	661.97 (29.40)	77.30 (10.76)	237.97 (12.31)
Superv. (half)	12.01 (1.52)	35.39 (1.57)	13.20 (1.84)	37.90 (1.96)
Random-5 (three draws, min–max)	7.96–8.10 (1.01–1.02)	22.70–23.04 (1.01–1.02)	7.20–7.29 (1.00–1.01)	19.35–19.60 (1.00–1.01)
ITI (48 heads, α=15)	9.75 (1.23)	31.76 (1.41)	8.32 (1.16)	24.25 (1.25)
ITI (24 heads, α=15)	8.81 (1.11)	27.05 (1.20)	7.76 (1.08)	22.24 (1.15)

G.11 What the Selected Features Read

For every behaviour, model, and dictionary, the steering scripts' rule (descending training $|\mathrm{AUROC} - 0.5|$ over the stored probe activations) was replayed to recover the selected feature indices, and each selected feature was scored on all 34 probe sets. Table 21 names, for each of the five top-ranked features, the probe set it separates best (largest $|\mathrm{AUROC} - 0.5|$) and how many of the five read the selecting concept best; Table 22 gives the highest-activating probe text of each rank-1 feature. The toxicity selections are built from features that read another probe set better than the toxicity set itself, usually a narrower attribute set and in three cases refusal or the code control (none of the 15 top-5 toxicity features at 1.4B and none of the five at 2.8B reads toxicity best), whereas the insult and threat selections mostly read their own concept (10 of 15 and 11 of 15 at 1.4B, 4 of 5 each at 2.8B), and some threat selections include a feature that reads the code control in the negative direction. Within a dictionary the toxicity and insult selections overlap (Jaccard 3/7 at seed 42 and at 2.8B, 2/3 at seeds 123 and 7), while threat shares at most one feature with either. Across dictionaries the rank-5 toxicity subspaces are not aligned: the smallest principal angle between two seeds' subspaces is 27–42° and the largest 85–88°. These are in-sample descriptions from mean-pooled activations of the probe sets; they establish neither a single meaning per feature, nor a causal role, nor generalisation beyond these texts.

Table 21: **What the five selected SAE features of each behaviour read best.** For each model, behaviour, and dictionary seed, the probe set on which each top-5 feature has the largest $|\mathrm{AUROC} - 0.5|$ among the 34 sets (AUROC in parentheses; values below 0.5 mean that the feature fires on the negative side), in selection order, and the number of the five that read the selecting concept best (Own).

Model	Behaviour	Seed	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Own
1.4B	toxicity	42	insult (0.97)	insult (0.90)	threat (0.95)	sexually explicit (0.96)	profanity (0.90)	0/5
1.4B	toxicity	123	insult (0.97)	sexually explicit (0.98)	refusal (0.92)	insult (0.77)	insult (0.88)	0/5
1.4B	toxicity	7	profanity (0.91)	profanity (0.94)	refusal (0.96)	sexually explicit (0.96)	insult (0.89)	0/5
1.4B	insult	42	insult (0.97)	insult (0.90)	insult (0.84)	insult (0.83)	profanity (0.90)	4/5
1.4B	insult	123	insult (0.97)	insult (0.88)	refusal (0.92)	insult (0.80)	insult (0.77)	4/5
1.4B	insult	7	insult (0.89)	profanity (0.94)	profanity (0.91)	insult (0.83)	refusal (0.96)	2/5
1.4B	threat	42	threat (0.95)	threat (0.87)	threat (0.84)	code (0.13)	threat (0.83)	4/5
1.4B	threat	123	threat (0.94)	threat (0.89)	refusal (0.92)	threat (0.82)	code (0.14)	3/5
1.4B	threat	7	threat (0.95)	threat (0.90)	threat (0.88)	code (0.06)	threat (0.78)	4/5
2.8B	toxicity	42	profanity (0.99)	sexually explicit (0.97)	insult (0.83)	code (0.00)	insult (0.78)	0/5
2.8B	insult	42	profanity (0.99)	insult (0.83)	insult (0.79)	insult (0.78)	insult (0.76)	4/5
2.8B	threat	42	threat (0.89)	threat (0.88)	code (0.13)	threat (0.86)	threat (0.81)	4/5

Table 22: **Highest-activating probe text of each rank-1 feature** (seed-42 dictionary of each model). Selection AUROC on the selecting concept, the probe set the feature reads best, the probe set the text comes from, and the first characters of the text; the texts are drawn from public toxicity resources and are offensive by construction.

Model	Behaviour	Feature	AUROC	Reads best	Source set	Text
1.4B	toxicity	7931	0.91	insult (0.97)	toxicity (earlier set)	The most shocking forbidden dirty dark incestuous pedophile secr…
1.4B	insult	7931	0.97	insult (0.97)	toxicity (earlier set)	The most shocking forbidden dirty dark incestuous pedophile secr…
1.4B	threat	13632	0.95	threat (0.95)	sexually explicit	"There were threats to mutilate my genitals, threats to slit my…
2.8B	toxicity	20350	0.88	profanity (0.99)	financial crime	fuck this shit
2.8B	insult	20350	0.98	profanity (0.99)	financial crime	fuck this shit
2.8B	threat	15797	0.89	threat (0.89)	hate speech	How do you break someone's arm?

H Changes from a Prior Submission

An earlier version of this work was reviewed at another venue. Reviewers identified, and we confirmed, defects that the present design corrects; this appendix records them and the controls added since. Feature selection and evaluation shared data, and the headline correlation reused one early-strength value per concept across its five features; we now enforce grouped, selection-disjoint folds with concept-level inference (§3.2, §3.4), report intervention effects as fixed- and refit-probe sensitivities against a random-subspace null over all 32 concepts (§5.5), and confine the early-strength predictor to Appendix G.9 with cluster-bootstrap intervals. The refusal probe's synthetic benign templates created a style confound, and its reported early emergence was a lexical artefact; negatives are now natural instructions, every claim is floor-corrected (§3.3), and a bootstrap re-analysis of that pipeline's selected features shows them flipping polarity across the first sixteen steps before saturating as a benign-style detector, confirming that the reported early emergence was the probe set, not the model; no claim about refusal behaviour or alignment remains. Cross-architecture claims rested on final-checkpoint matching plus a raw-neuron trajectory; we now train identical SAEs on OLMo-2 intermediate checkpoints (§5.4) and drop the final-checkpoint matching. Single-seed SAEs left feature-identity uncertainty unmeasured; three seeds now bound it, and the dictionaries' token budget, previously confounded with the checkpoint, is equalised in a pre-registered control (Appendix F.1). Generation-level evaluation reused selection prompts and the selection classifier; prompts are now disjoint, the scorer is independent, a second scorer replays every certificate, and pre-registered confirmatory rounds replicate the design on fresh prompts, two further behaviours, and a second model family (§5.5, Appendix G). An exploratory control run after those rounds removes dictionary subspaces selected for unrelated concepts and shows that the random-direction null understates the effect of removing active dictionary directions (Appendix G.4). The comparison set was one supervised direction; it now includes linear probes as the primary readout and mean-difference, INLP, LEACE, activation-addition, RepE, ITI, and feature-clamping edits, with held-out perplexity reported for each (Appendix G.10). Scale was limited to 1.4B-class models; the reading-side results now extend to Pythia-12B (§5.4). A name that implied a new architecture ("concept-specific SAE") is dropped: the dictionaries are standard TopK SAEs, and the contribution is the audit protocol. Compute claims lacked a ledger; every run now logs one (Appendix F).

Two statements in the earlier text misreported their own numbers; the reported numbers are regenerated from the released artifacts and bound to their source files by an automated check. Auditing the redesigned pipeline post hoc, we found two further defects. First, all readout folds were initially split at row level; on probe sets built from minimal pairs or from texts that share a question this placed an example's oppositely-labelled mate in training, which can invert probe generalisation (Theorem 1), and it produced an apparent SAE reading advantage on five bias axes, of which grouped folds remove all but gender (§3.2, §5.1). Second, batched generation used right padding while the generation utility ignores attention masks, decoupling continuations from their prompts; all generation results now come from unbatched decoding, and the earlier generation-level numbers are void.