

Training-Free Prefill Safety Blocking via Cross-Layer Interpretability

Madhavi Evana
Amazon Web Services
imevana@amazon.com

Yiwen Zhang
Amazon Web Services
yiwenzhg@amazon.com

Abstract

A neural network often settles on its answer long before it shows it. We make that precise with a training-free tool, the sign lens. It reads a concept straight from a hidden state using only the sign of each coordinate, positive or negative (Nalagatla, 2026): a concept is a set of coordinates that share a fixed sign pattern, detected by counting agreements. No probe, no dictionary, no training.

With this lens we follow one concept through the network, and a sharp pattern in time appears. The answer is fixed early, within the first tenth of the depth, by the computation that reads the input. But it stays hidden there: it cannot be read out until the last third, when the network finally surfaces it. The concept is written early and read late, and the two moments are far apart.

This gap is not a quirk of one model. We find it in every model we test, from small to very large, across five kinds of architecture, including ones with no attention at all, and across text, vision, and speech. It reflects how deep networks compute, not anything special about the transformer. The sign we read is also the causal one: suppress it and the behavior changes, while a matched control does nothing.

The account earns its keep. Because a network commits to a harmful request early, before it emits a single token, we can read that commitment at the prompt and refuse to answer. This training-free input guardrail matches a supervised probe on detection while using none of its own labels, though a dedicated supervised guard remains more precise on the hardest cases. We are equally clear about a limit: a sign is easy to read and hard to rewrite cleanly. Interpretability usually asks what each part of a network encodes. We ask a sharper question: when is a feature decided, when can it be seen, and when can it be changed?

1 Introduction

To interpret a network, the usual move is to train a second model. Sparse autoencoders fit an overcomplete dictionary to its activations (Cunningham et al., 2023; Bricken et al., 2023; Templeton et al., 2024); linear probes are trained per concept on labelled data (Alain & Bengio, 2017; Belinkov, 2022). Both work, but both cost compute and curated data, and both carry a subtler risk: what they recover is a property of the auxiliary objective, not necessarily of the model. A recent survey puts it plainly (Sharkey et al., 2025). These methods are data-hungry, leave the geometry of features unexplained, struggle with structure spread across heads and layers, and describe a network’s activations rather than the mechanisms that produce them.

Our tool is simpler. The sign lens reads content straight off the standard basis (Nalagatla, 2026), in the spirit of the logit lens (nostalgabraist, 2020). The sign of each coordinate is content; its magnitude is confidence; a concept is a set of coordinates that share a sign pattern, read by counting agreements (Section 3). Unlike the logit lens, it applies to the attention K/V and the MLP write-space, not just the residual stream. Unlike SAEs, probes,

or the tuned lens (Belrose et al., 2023), it needs no dictionary, no rotation, and no training. We use it to ask three questions about a concept at once, and to show their answers form a single lifecycle: where it lives, when it appears, and whether it drives behavior.

Contributions. We make four contributions, organized around that lifecycle. First, on the statics of where a concept lives (Section 4), we show that coalitions are readable well beyond the residual stream, in attention K/V and, most sharply, in the MLP gate activation (per-dimension AUC ≈ 0.93), so the sign is legible across the computation graph and not only in the residual. This holds from 0.5B to 32B parameters and across text, vision, and audio; the sign is the causal variable; and the reading is competitive with a trained SAE and a supervised probe. Second, on the dynamics of when a concept appears (Section 5), we show that a context-computed answer becomes readable only late (mean $\approx 71\%$ of depth), at a model-specific depth that behaves as an architectural fingerprint, while lexical properties are readable from the embedding and never assemble. This holds across five architecture classes (dense, mixture-of-experts, state-space with no attention, Griffin, and RNN) and across text, vision, and speech, so it is not a transformer or attention artifact; the answer is present early and extracted late, and even a no-forward-pass baseline that intersects the input tokens’ unembedding signs recovers it at median rank 12 of 152,064. Third, on the write→read gap (Section 6), causal tracing localizes the write early (subject-enrichment at $\approx 10\%$ of depth) while the sign lens reads the answer only late ($\approx 70\%$): a ≈ 60 -point-of-depth gap that is unanimous across all 11 models and is corroborated by steering-induction; and reading a concept is not the same as being able to change it, since the coalition is causally manipulable from early layers, long before it is readable. Fourth, we assemble these into a lifecycle map and turn it into training-free tools (Section 7, Section 8): a subspace-fused input guardrail (harmful-vs-benign AUC 0.94–0.97 across three scales, matching a supervised probe on detection with no training and halting before the first generated token; a dedicated supervised guard is more precise on hard negatives, Section 10) and a correctness monitor (AUC up to 0.94).

Implications. For the field, this shifts interpretability from a static dictionary of what each dimension encodes to a temporal circuit of when each feature is written, readable, and decisive: a parts list becomes a schematic. It also explains a standing puzzle: mid-forward-pass sign flips were often inert because the answer is written early and extracted downstream. For the mechanism, we obtain the full (subspace $\times$ depth) map of a concept: written early (the FFN induction band, ≈ 25 – 40% depth on Gemma-3-4B via steering induction; Section 6 gives the ROME subject-enrichment estimate of $\approx 11\%$ across all 11 models), readable at parity across the residual and the MLP write-space (gate_act) in the late band, and causally decisive only near the output. For deployment, this is the step from “we can read features” to “we can read them early enough, and in enough places, to act before the model commits”: the guardrail halts a harmful request at prefill, stays robust because the concept is readable in more than one subspace, and applies any intervention at the causally decisive layer rather than the merely readable one, with no trained probe, dictionary, or auxiliary model.

2 Related work

Training-free lenses and standard-basis reading. The logit lens (nostalgebraist, 2020) and its trained refinement the tuned lens (Belrose et al., 2023) decode intermediate residual states, but only in the residual stream and, for the tuned lens, with a learned per-layer probe. The standard basis can also be read directly, with each dimension’s sign encoding content and its magnitude confidence, giving a training-free feature basis at a single layer (Nalagatla, 2026). Our sign lens reads the standard basis across depth, resolving when a concept is written, becomes readable, and is decisive.

Sparse dictionary learning and its critique. Sparse autoencoders decompose activations into overcomplete, interpretable dictionaries (Cunningham et al., 2023; Bricken et al., 2023; Templeton et al., 2024), but recent work questions whether their features beat simple baselines under probing (Kantamneni et al., 2025) or recover ground-truth structure (Anonymous,

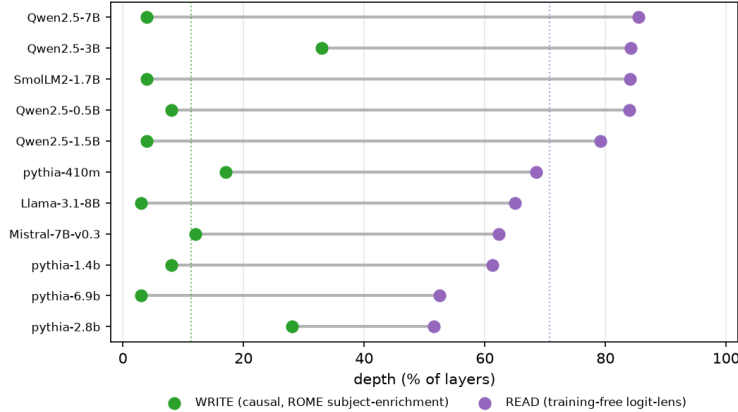


Figure 1: The phenomenon, in one figure. For each of 11 models, the causal write locus (green; ROME subject-enrichment depth) and the training-free read locus (purple; logit-lens depth). A context-computed answer is written at $\approx 11\%$ of depth on average but becomes readable only at $\approx 71\%$: a ≈ 60 -point write $\rightarrow$ read gap, on every model. Reading is the last stage to arrive, not the first.

2026). This debate motivates a training-free reader; we reproduce SAE-detection parity ourselves against Gemma Scope 2 (Section 4) rather than only citing it.

Probes, directions, and steering. Linear probes read concepts from activations but require labels and training (Alain & Bengio, 2017; Belinkov, 2022) and, being correlational, cannot by themselves certify a causal variable. The linear representation hypothesis (Park et al., 2024) and representation-engineering and refusal-direction work (Zou et al., 2023; Arditì et al., 2024) show that a few directions can steer behavior; we locate when in depth that leverage exists, and find (Section 10) that reading is reliable while surgical rewriting is not.

Where facts live and when they are extracted. Causal tracing localizes stored facts to mid-layer MLPs at the subject position (Meng et al., 2022), with subject enrichment early and attribute extraction to the last position late (Geva et al., 2023; Yu et al., 2024); the pattern is echoed across architectures (Zhang et al., 2025) and by attribution-graph tracing that reports late, planned answer assembly (Anthropic Interpretability Team, 2025a;b). Our write $\rightarrow$ read gap quantifies this early-write/late-read separation with a training-free reader on 11 models.

Concept formation across depth. Closest is concurrent work treating concept formation as depth-extended rather than a single-layer event: the Concept Allocation Zone (Henry, 2026a) and its companion (Henry, 2026b), and cross-layer discrete concept discovery (Anonymous, 2025). These characterize where concepts allocate (via continuous separation geometry) or learn a decomposition; we instead read the standard basis discretely, study externally-verifiable context-computed answers, and establish a causal lifecycle (present-before-extracted and read $\neq$ cause) that those works do not.

3 The sign coalition (statics)

A concept is a coalition of standard-basis dimensions with a fixed sign pattern: content lives in the sign, and confidence in the magnitude. Concretely, for a hidden state h we take the sign vector $s = \mathbf{1}[h > 0]$; a concept c is identified with a set of dimensions $\mathcal{D}_c$ together with a polarity π , and its score is the fraction of coalition dimensions whose sign matches π . No dictionary, rotation, or trained parameters enter the reading. The treatment of dimensions as independent registers is empirically justified: pairwise mutual information between dimensions is below 0.006 bits, so the “bag” abstraction is not merely convenient but

reflects near-independence of the standard-basis coordinates (Nalagatla, 2026), consistent with superposition placing distinct features on approximately axis-aligned directions (Elhage et al., 2022; Park et al., 2024).

Two method details carry through the rest of the paper. First, the unsupervised sparsity claim uses mean-centered signs $\text{sign}(h - \mu)$; supervised detection does not require centering. Second, all single-token reads use BOS-prefixed caches: a bare single token has no attention context, and the prefix is required to obtain clean states. We confirm this matters for the write-space reads specifically: adding the prefix sharpens residual retrieval AUC from 0.73 to 0.89 and gate-activation retrieval from 0.62 to 0.87, and we adopt it as the common convention throughout.

4 Where a concept lives: subspaces and modalities

Coalitions are sign-readable well beyond the residual stream, and the sign is the causal variable rather than an epiphenomenon of magnitude. A 175-category standard-basis survey (BOS-prefixed, post-RoPE keys) that we build on (Nalagatla, 2026) finds coalitions readable in the residual (175/175 categories at AUC 0.86), in the attention keys (175/175 at 0.90) and values (173/175 at 0.87), and, most sharply, in the MLP gate activation (175/175 at 0.93), consistent with FFN neurons writing axis-aligned content (Geva et al., 2023). A full (subspace $\times$ depth) surface (Fig. 4) confirms that the residual and gate_act are strongly sign-readable across depth (peak AUC ≈ 0.88 each on Qwen2.5-1.5B); a lightweight K/V probe reads keys pre-RoPE and is correspondingly weaker (0.59/0.63), so the strong K/V claim rests on the post-RoPE survey above. The structure is not language-specific: the same per-dimension sign organization appears in vision (DINOv2, ViT) and audio (AST) encoders, with no BOS analog needed.

The sign is causal. Negating a coalition drops the target logit substantially, by 9–16 on Qwen2.5-7B and by 17–25 on Gemma-3-4B across depth, while a “toward” intervention and a matched disjoint-concept control sit near zero. And the reading is competitive with trained methods that we run ourselves. On harmful-vs-benign detection the training-free sign lens matches a supervised probe (AUC 0.94–0.97 across three model scales) with no training, using only a small set of anchor words (Fig. 2). Against a trained SAE, we read Gemma-3-4B’s residual stream with the Gemma Scope 2 JumpReLU dictionaries (Lieberum et al., 2024) and score nine concepts by 5-fold cross-validated detection AUC: the training-free sign coalition matches or exceeds the best per-concept SAE latent on 6–7 of nine concepts (mean sign AUC 0.94/0.96 at layers 17/22, vs. 0.88/0.95 for the best SAE latent), approaching a supervised probe (0.997) at zero training cost, and the same parity holds on Gemma-3-1B, consistent with reports that SAE features need not beat simple baselines on real tasks (Kantamneni et al., 2025; Anonymous, 2026). Unlike the logit lens, which is also training-free, the sign lens applies to attention K/V and the MLP write-space, not just the residual.

We are explicit about provenance. The 175-category post-RoPE subspace survey above is from the companion coalition study (Nalagatla, 2026), which is publicly available and can be checked directly; everything else in this paper is established here: the cross-layer dynamics (Section 5), the write→read gap (Section 6), the cross-architecture and cross-modal generalization, the SAE- and probe-parity reproduction on Gemma above, and the guardrail and correctness monitors (Section 8). Our own subspace reproduction reads the residual and gate_act sharply but keys and values only pre-RoPE (the weaker 0.59/0.63), so the strong post-RoPE K/V number is one we cite rather than reproduce.

5 When a concept becomes readable: cross-layer dynamics

Context-computed answers are present early and read late, whereas lexical properties do not assemble at all. As a control, lexical categories (animal, color, fruit, body-part) are readable from layer 0 at AUC ≈ 0.9 –1.0 and stay flat across depth, while a random-concept null sits at chance: there is nothing to “assemble” for a property that is already in the embedding. A context-computed answer behaves oppositely, emerging only late, and this late emergence is architecture-general, not attention-specific. Tracking the answer with the sign metric (which

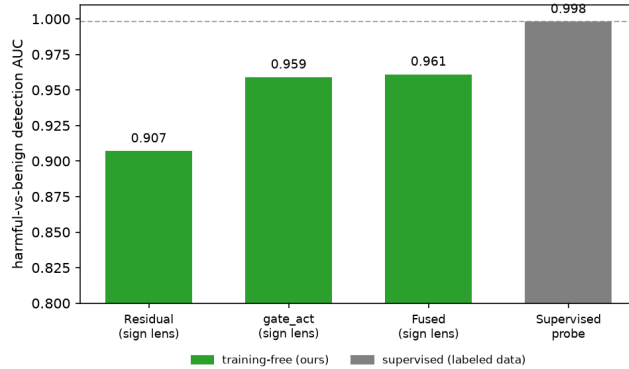


Figure 2: Why the sign lens. Harmful-vs-benign detection AUC (Qwen2.5-1.5B): the training-free sign lens (green) matches a supervised probe on the same states, with no training and only a small anchor lexicon. An SAE reaches parity but needs dictionary training; the logit lens is training-free but residual-only. The sign lens is the only option that is training-free and reads beyond the residual stream.

tracks the logit lens) across five architecture classes with at least two models each, we find late extraction in dense, mixture-of-experts, state-space (with no attention), Griffin-hybrid, and RNN models (Table 1). A model with zero attention (Falcon-Mamba) shows the same late-extraction signature, so the phenomenon is a property of depth-structured next-token computation, not of the transformer or of attention; this parallels cross-architecture analyses of factual recall (Zhang et al., 2025).

The answer is present early and merely extracted late. A no-forward-pass baseline that simply intersects the input tokens’ unembedding signs already recovers the answer at median rank 12 of 152,064 (top-1 9% on Qwen2.5-7B), so part of the answer is lexically present in the prompt before any computation. This lexical prior is real but limited, and forward computation does far more than echo it: a readout of the residual improves top-1 from that 9% to $\approx 99\%$ (logit lens) and 67% (a trained decoder) in the late band. Crucially, the improvement is late. At the early write locus ($\approx 11\%$ depth) the same decoder reads the answer at only $\approx 1\%$ top-1, below the no-forward-pass baseline, so early layers causally set the answer without yet making it decodable. The lexical prior and the late decodability are distinct facts, and neither collapses the write $\rightarrow$ read gap: the late layers extract rather than form the answer. The exact readout depth behaves as a family-linked fingerprint. A one-way permutation ANOVA over the 12 dense models (6 families) finds a significant family effect ($F=8.5$, $p=0.008$, 20k permutations; removing the smallest Pythia, a size outlier, strengthens it to $p=0.0009$); family means span $\approx 59\text{--}84\%$ of depth, whereas scale does not move it (Qwen2.5 reads within ≈ 7 points from 0.5B to 7B) and instruction tuning leaves it unchanged. Within-family spread (about 7 points for Qwen2.5, 16 for Pythia) is smaller than the across-family span, so family is a better predictor of read depth than scale or any single structural factor (parameter count, layer count, attention, mixture-of-experts, or recurrence). With only two to four models in most families we report this as a predictive fingerprint, not an irreducible law. Finally, the read is not confined to the residual: across 6 models and 3 families (Qwen2.5-1.5/3/7B, Mistral-7B, Pythia-1.4/6.9B) the MLP write-space (gate_act) reads the context-computed answer at parity-or-later than the residual, never earlier, at peak AUC up to 0.90 (Fig. 3). This supports subspace universality (the answer is readable in the write-space too) but explicitly rules out an earlier-warning claim; the write-space is a redundant read site, which is precisely what makes the fused monitor of Section 8 robust. The full (subspace $\times$ depth) surface, including K/V, is in Fig. 4.

Beyond text: vision and speech. The same late-extraction signature holds in other modalities, read with the same sign lens. In vision, the predicted class becomes readable late in ViT-L (at $\approx 71\%$ depth, sign margin $+0.38\text{--}0.48$) and in attention-free convnets (ConvNeXt, ResNet-50), where the class-separating sign margin rises monotonically to a late peak. In

Architecture class	Models (≥ 2 each)	Late-emergence depth	Concept-specific causal
Dense (transformer)	12 models (Qwen2.5, Pythia, ...)	52–85% (mean 71%)	own $\gg$ disjoint, 8–10/12
Mixture-of-experts	OLMoE-1B-7B, Granite-3B	75–88%	−1.14 vs −0.84, $p\approx 0$
State-space (no attention)	Falcon-Mamba-7B, Mamba-2.8b	67–84%	− 4.70 vs −0.50, $p\approx 0$
Griffin-hybrid	RecurrentGemma-9B/2B	90–100%	−2.51 vs −1.31, $p=0.04$
RNN / linear-attn	RWKV-4 3B/1B5	79–97%	−1.24 vs −0.79, $p=0.02$

Table 1: Read-late and concept-specific causation are architecture-general. All five classes, including one with no attention (state-space) where the effect is in fact strongest, show the context-computed answer emerging late and a concept-specific sign-flip (own coalition $\gg$ a disjoint concept’s). Causal load-bearing vs. a magnitude-matched null holds on every model; own $\gg$ disjoint concept-specificity holds in 6/8 second-model tests (reported as a bound, not a universal law).

speech, CTC encoders (wav2vec2, HuBERT) extract the character at $\approx 67\%$ depth, while the Whisper decoder holds the token later ($\geq 92\%$), a depth fingerprint within a modality. Late extraction is thus not a language- or attention-specific artifact but a property of deep next-token (or next-label) computation across text, vision, and speech, with and without attention.

6 The write $\rightarrow$ read gap: when a concept causes behavior

A concept is written (causally) early but becomes readable late, and the two loci are separated by a large, model-general depth gap. We establish this two ways that agree.

The first is causal tracing on our 11 models (Fig. 1). ROME-style restoration (Meng et al., 2022) localizes the write: restoring the clean state at the subject position recovers the answer from early layers, with subject-enrichment peaking at mean $\approx 11\%$ of depth (range 3–33% across Qwen2.5 (Qwen Team, 2024), Pythia (Biderman et al., 2023), Llama-3.1, Mistral, and SmolLM2), whereas restoring the last position recovers it only late (peak $\approx 98\%$ of depth, unanimous). The training-free logit-lens read of the answer emerges at mean $\approx 71\%$ depth. So the answer is causally written roughly 60 points of depth before it can be read out, on every model tested. Both loci are measured on the same models and the same factual-recall battery; they differ in two intrinsic ways: the probe (causal state-restoration for the write, decodability for the read) and the token position each reads (the subject position for the write, the last position for the read), the latter being the standard subject-enrichment-then-extraction structure of recall (Geva et al., 2023). Neither difference is an artifact of comparing different prompts or model sets: the same-prompt, same-model comparison is exactly what a lifecycle measurement requires. What is not yet closed is a same-method write number across families, which the steering run below supplies for one family and we flag as a pending multi-family cross-check.

The second is steering induction, a complementary and independent probe: steering a concept coalition raises its target logits from early layers (Gemma-3-4B, layers 8–13, or ≈ 25 –40% depth), the generative-write locus. A naive single-layer reading instead peaks at a deep intervention-sensitivity layer, not the write location. Two different interventions reach the same conclusion, that the concept is written early. They are applied to non-overlapping models, however (ROME across the 11-model suite, none of them Gemma; steering on Gemma-3-4B), so a same-model comparison of the two write probes is part of the pending cross-check. This early-write, late-last-position-extraction pattern matches independent accounts of factual recall (Geva et al., 2023; Yu et al., 2024) and attribution-graph tracing that reports late, planned answer assembly (Anthropic Interpretability Team, 2025a;b).

We are precise about what the write locus measures. ROME restoration localizes where a state is causally necessary: reinstating the clean subject-position representation early recovers the answer. This is a proxy for “written,” not a direct decode of the answer at the subject position, and it is compatible with the subject representation being reused downstream rather than the answer being fully set at 11%. The honest statement of the gap is therefore between causally in place early and decodable late. A same-method decode

at the subject position across depth, together with the multi-family steering cross-check we flag as pending, would turn this multi-probe convergence into a single-probe measurement; we state the gap as strong convergent evidence and no stronger.

Reading a concept is not the same as being able to change it. Negating the answer coalition changes the output from early layers: on Qwen2.5-7B the answer logit drops by 9–16 across the whole stack (block 0: -16.5 ; block 8: -9.1) versus ≈ -1 for a random flip of equal size, even though the answer only becomes readable at $\approx 85\%$ depth. So the coalition is causally manipulable well before it is decodable: reading is late, but acting is not. A stricter concept-specific flip (the concept’s own coalition versus a disjoint-concept control of equal size) holds in 8 of 12 dense models at the fixed read-layer (FDR-controlled), and in 10 of 12 when each concept is evaluated at its own peak-effect layer. The four fixed-layer exceptions show no size or family pattern (within Qwen2.5, the 0.5B and 7B models are significant while 1.5B and 3B are not), and are either marginal near-misses ($p = 0.06, 0.11$) or non-directional at that single layer, so the fixed-read-layer test is conservative rather than a model-class failure. We report concept-specificity as a bound, not a universal law. This concept-specific causation generalizes across architecture classes (Table 1): $\text{own} \gg \text{disjoint}$ holds for mixture-of-experts, Griffin, and RNN models, and, most strongly, for the attention-free state-space model (Falcon-Mamba, own -4.70 vs. disjoint -0.50 , $p \approx 0$, beating both a disjoint concept and a random null). The sign is thus the causal variable regardless of whether the architecture even has attention. This picture is consistent with representation-engineering and refusal-direction findings that a small number of directions can steer behavior (Zou et al., 2023; Arditì et al., 2024), but it localizes when in depth that leverage exists.

7 The lifecycle map

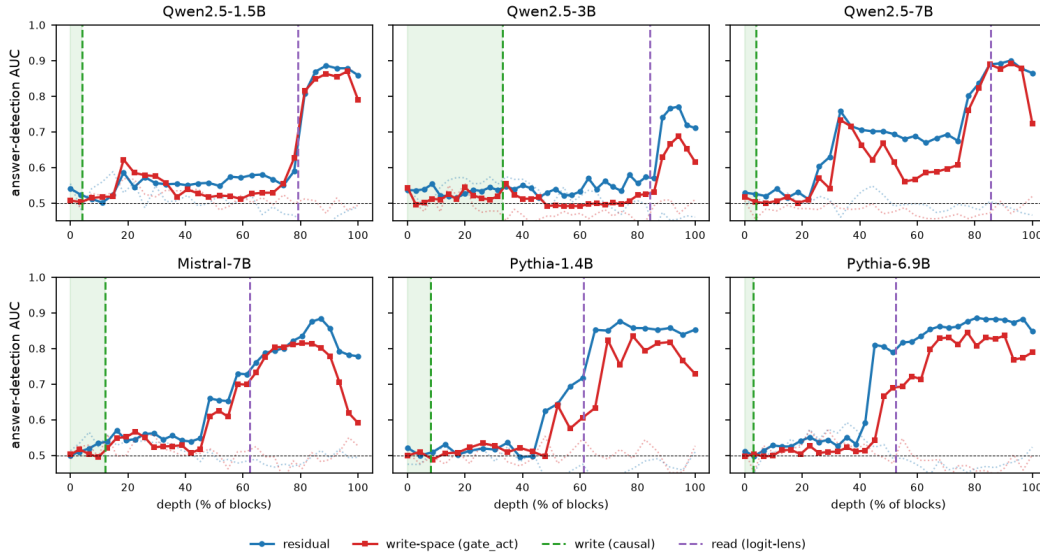


Figure 3: The lifecycle of a feature (6 models / 3 families, BOS-aligned). Per model: answer-detection AUC across depth in the residual vs. the MLP write-space (`gate_act`), shuffled-answer null dotted; the green line marks the write locus (ROME subject-enrichment depth) and the purple line the read locus (logit-lens depth). The concept is written early (green) but the sign-read curves rise, and the answer becomes readable, only late (purple): a large write→read gap on every model. The write-space tracks the residual at parity (redundant read sites, not an earlier channel).

Putting the three axes together yields a single trajectory. A concept is written early (in the FFN, with subject-enrichment at $\approx 10\%$ depth), becomes readable late and at parity across the residual and the MLP write-space (at $\approx 70\%$), and is causally manipulable across the

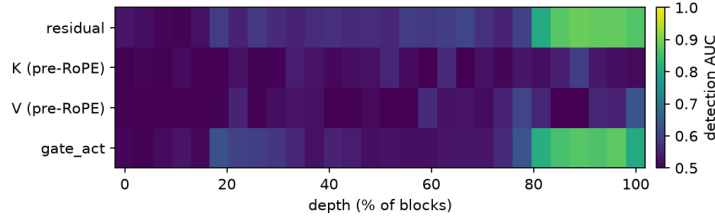


Figure 4: The (subspace $\times$ depth) surface (Qwen2.5-1.5B). Answer-detection AUC of the context-computed answer read with the sign lens in the residual, attention K/V, and the MLP gate_act, per block. Residual and gate_act are strongly readable and peak late; K/V here are read pre-RoPE and are weaker (post-RoPE K/V reach ≈ 0.90). The surface makes concrete that the concept is readable across the computation graph, not just the residual.

stack from early layers. Reading is therefore the last of the three stages to arrive, not the first: the schematic that a static feature catalogue cannot provide.

8 Applications: the payoff

The lifecycle account is directly operational, and its two clearest payoffs need no training: an input guardrail and a correctness monitor.

The first is a training-free input guardrail. Because the harm concept is present and readable in the prompt at prefill (Section 5), we read the harm coalition at the last prompt token and, if it fires, halt before the first token is generated, so the model never emits the harmful content at all. This is the defining difference from an external guard such as Llama Guard, a separate model that runs its own inference pass: our check instead reads the target model’s own interior state at the input, at prefill, and blocks emission before the first token is generated, with no second model in the loop. The harm coalition is built training-free: from a small set of harm anchor words (bomb, weapon, poison, malware, kill, ...) and a diverse neutral background set, we keep the dimensions whose sign separates the two, together with the harm-side polarity, using no gradients, dictionary, or trained probe beyond that small anchor lexicon. A prompt is then scored at prefill by the fraction of those dimensions whose sign, at the last prompt token, matches the harm polarity (fused across the residual and the MLP write-space), and crossing a threshold triggers the halt. Because this reads the model’s own internal representation of the request rather than its surface words, it separates a genuinely harmful request from a benign one that merely reuses a charged word (e.g. “kill a runaway process”). On 300 AdvBench harmful versus 300 alpaca benign prompts, with a 76-prompt hard-negative set of benign-but-harm-adjacent requests (e.g. “kill a runaway process in Linux”, “detonate a controlled demolition scene in my video game”), the training-free fused sign gate reaches harmful-vs-benign AUC 0.94–0.97 across Qwen2.5-1.5B, 3B, and 7B, approaching a supervised logistic probe on the same states (AUC 0.998–1.000, 5-fold CV). The point is not to beat the probe but to match its detection without any training (only a small anchor set) while remaining mechanistically interpretable. Fusing the residual and the MLP write-space is the robustness win the parity result licenses: which single subspace is cleanest shifts with scale, as the write-space carries the signal at an early layer in the smaller models, the residual at a late layer at 7B, so reading both and fusing keeps the monitor robust to any one site being noisy, and hard-negative false positives fall with scale (0.62, 0.28, 0.07 at 1.5B/3B/7B, at 90% recall). We are explicit about the ceiling: the supervised probe is strictly better on hard negatives (zero false positives at every scale), so the training-free gate matches detection, not hard-negative precision (Section 10). Deployment cost is near-zero: the monitor reads existing hidden states from the prefill pass the model already runs and thresholds a sign-agreement count, generating no extra tokens, whereas an external guard model such as Llama Guard (Inan et al., 2023) is a separate model that requires its own inference pass, adding latency and deployment cost. The safeguard is thus effectively free at the point of the request.

The second is a correctness monitor: the late-band sign-commitment margin predicts whether the model’s own answer is correct from a training-free, fixed penultimate-layer read. The per-model AUC tracks the model’s own competence rather than varying at random: it runs from 0.65 on the weakest models (Pythia-1.4B, base accuracy 0.69) to 0.94 on the strongest (Mistral-7B and Gemma-2-9B, base accuracy ≈ 0.97), mean 0.78 over 12 dense models (Table 2). On the same task the sign margin and a plain logit-lens confidence read are comparable: the sign margin leads on 7 of 12 models and the logit lens on 4 (one tie), with mean AUCs within 0.01, so the monitor’s value is being training-free and fixed-layer, not beating the lens. It too is architecture-general, reaching AUC 0.63–0.85 across mixture-of-experts, Griffin, RNN, and the attention-free state-space model (all above chance, in the same range as dense), with a best read layer that tracks each architecture’s extraction fingerprint. Finally, because the coalition is causally manipulable before it is readable (Section 6), a controller can act at prefill or early layers rather than waiting for the readable band: read $\neq$ cause has a direct consequence for where to steer or halt.

9 Discussion: open problems addressed

Our results speak to several problems the field has named as open (Sharkey et al., 2025), and it is worth being concrete about which result addresses each, and how far. The first is a decomposition that does not depend on training or curated data. The sign lens (Section 3) meets this directly: it reads a concept with a fixed rule and no learned parameters. The second is explaining the geometry of features rather than treating them as an unstructured bag. We show the standard-basis coordinates behave as near-independent registers, and we read coalitions that span heads, layers, and subspaces (Sections 3 and 4). The third is extending the logit lens beyond the residual stream; the sign lens does this by reading the attention K/V and the MLP write-space directly (Section 4). The fourth, which we take furthest, is relating a network’s activations to the computation that produces them. By localizing where a concept is causally written and separating that from where it becomes readable (Section 6), we tie a readable feature to the moment the mechanism sets it. The last is real-time monitoring, which we provide as the training-free input guardrail and the correctness monitor (Section 8). We address these in part, not in full: we localize where a concept is written without reconstructing the full circuit that writes it, and our reading identifies a causal variable without giving a complete causal model.

10 Limitations and threats to validity

We state the regimes where the sign lens holds and where it degrades rather than averaging over them. The read is strongest in the residual and the MLP write-space, for models with a healthy dimension-to-vocabulary ratio ($\gtrsim 1.5B$ parameters), and for lexical or valence concepts; it degrades for very small models (Qwen2.5-0.5B, low D/V) and for keys read pre-RoPE, and its concept-specific causal effect holds in 8 of 12 dense models at the fixed read-layer (10 of 12 at each concept’s peak layer), not universally.

Several specific caveats follow. The write $\rightarrow$ read gap is quantified by causal tracing on all 11 models (our own runs); a same-method steering-induction number across families is a deliberate pending cross-check, since reproducing it faithfully needs the companion paper’s gradient-based steering pipeline. We therefore corroborate the write locus with steering on one family (Gemma-3-4B) and carry the multi-model result with causal tracing, and we name this as the one open item on the paper’s central claim rather than leave it implicit. The write-space reads at parity with the residual, not earlier, so we make no earlier-warning claim, and we report the non-significant concept-specific flips as a bound rather than a universal law. The guardrail is evaluated on AdvBench and alpaca (300 prompts per class) across three model scales (1.5B, 3B, 7B) with an expanded, deliberately adversarial 76-prompt hard-negative set (benign requests built from harm vocabulary); a trained probe remains strictly better on those hard negatives (zero false positives at every scale), so the training-free gate matches detection AUC, not hard-negative precision. Because these negatives are adversarial by construction, the 1.5B false-positive rate is a stress-test ceiling rather than a typical operating point, and it falls sharply with scale. As an external output-guard baseline

we run Llama Guard 3 8B on the same prompts. It is a strong supervised guard and is substantially better than our training-free gate on hard-negative precision: hard-negative false-positive rate 0.01, versus our 0.62/0.28/0.07 at 1.5B/3B/7B (AUC 0.999, recall 0.98). We do not claim to match it on precision. The training-free gate instead wins on a different axis: it needs no labels or training, reads the target model’s own prefill states at near-zero added cost, and halts before the first token, whereas Llama Guard is a separate 8B model that requires its own inference pass. The two are complementary, not substitutes. The smaller Llama Guard 3 1B was unreliable in our single-turn harness, marking overtly harmful prompts as safe, so we do not rely on it. The only trained component anywhere is a ridge probe used as a control and the supervised probe baseline; “training-free” refers to the sign-lens method itself. Detecting a named concept still requires a small set of anchor words that specify it (a lightweight, human-provided form of supervision), but no labeled training corpus and no gradient training, which is the precise sense in which the reader is training-free; it is not label-free. Coalition dimensions are layer- and space-specific (cross-layer Jaccard ≈ 0.042), and the unsupervised-sparsity claim requires mean-centering, though supervised detection does not.

Two honesty points bound the contribution. First, the sign lens equals a probe rather than exceeding one: under both a detection test and a magnitude-matched interchange test, the training-free reader matches a trained linear probe and is not more causal, so the claim is probe-equivalent reading at zero training and labeling cost, not superiority. Second, the timing is descriptive rather than a free lunch: “extracted late” limits naive early exit (we measure $\approx 0\%$ layer savings at acceptable accuracy on dense models); the correctness monitor does not transfer to open-domain QA (on a small web-questions probe the late-layer sign read scores AUC ≈ 0.56 and the logit-lens confidence read ≈ 0.50 , both near chance, so neither the sign margin nor the lens is usable there); and sign-coalition flipping is not a clean surgical editor (its effect is model-dependent with comparable collateral). We report these as honest negatives: the contribution is the universal characterization plus the prefill safety and correctness monitors, not a speedup or a general-purpose editing tool. Relatedly, concept-specificity (own $\gg$ disjoint) holds in most but not all second-model tests (it fails on 2 of 8), while causal load-bearing versus a magnitude-matched null holds on all.

11 Conclusion

A concept in a transformer is not a static address in a basis; it has a lifecycle. Read with the training-free sign lens, a context-computed answer is causally written early ($\approx 11\%$ of depth), becomes readable only late ($\approx 71\%$), and is causally manipulable across the stack from early layers, a write $\rightarrow$ read gap that is unanimous across 11 models and readable in every subspace we test. Turning this temporal account into practice needs no training: a subspace-fused input guardrail matches a supervised probe on detection with no labels, halts before the first token, and adds near-zero overhead, and a late-band read predicts correctness before the output. Interpretability moves from a static catalogue of what each dimension encodes to a temporal circuit of when each feature is written, read, and decisive, and that timing is what makes it operational.

Reproducibility statement

All experiments use publicly available models (the families listed in Section 5 and Section 6) and datasets (AdvBench and alpaca for the guardrail; Gemma Scope 2 for the SAE comparison). The sign lens requires no training: a concept is specified by a small anchor word list and read by a fixed sign-agreement rule. We report detection thresholds and operating points with each result, and will release the code through an anonymized repository during the review period, together with the harm-anchor and background word lists and the per-concept anchor sets used in Section 4.

References

- Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In ICLR Workshop, 2017. arXiv:1610.01644.
- Anonymous. Cross-layer discrete concept discovery for interpreting language models. arXiv preprint arXiv:2506.20040, 2025.
- Anonymous. Sanity checks for sparse autoencoders: Do SAEs beat random baselines? arXiv preprint arXiv:2602.14111, 2026.
- Anthropic Interpretability Team. On the biology of a large language model. Transformer Circuits Thread, 2025a.
- Anthropic Interpretability Team. Circuit tracing: Revealing computational graphs in language models. Transformer Circuits Thread, 2025b.
- 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 NeurIPS, 2024.
- Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. Computational Linguistics, 2022. arXiv:2102.12452.
- Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. arXiv preprint arXiv:2303.08112, 2023.
- Stella Biderman, Hailey Schoelkopf, Quentin Anthony, et al. Pythia: A suite for analyzing large language models across training and scaling. In ICML, 2023. arXiv:2304.01373.
- Trenton Bricken, Adly Templeton, Joshua Batson, et al. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023.
- Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. arXiv preprint arXiv:2309.08600, 2023.
- Nelson Elhage et al. Toy models of superposition. Transformer Circuits Thread, 2022.
- Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In EMNLP, 2023.
- J. Henry. Tracking how concepts form across transformer depth (the concept allocation zone). arXiv preprint arXiv:2605.24856, 2026a.
- J. Henry. Geometric evolution maps: Extracting stable concept probes from transformer residual streams. arXiv preprint arXiv:2605.25848, 2026b.
- Hakan Inan et al. Llama guard: Llm-based input-output safeguard for human-ai conversations, 2023. arXiv:2312.06674.
- Subhash Kantamneni et al. Are sparse autoencoders useful? a case study in sparse probing. In ICML, 2025.
- Tom Lieberum, Senthooan Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on Gemma 2. arXiv preprint arXiv:2408.05147, 2024. Gemma Scope 2 extends these JumpReLU SAEs to the Gemma 3 family.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In NeurIPS, 2022.

- Varun Nalagatla. Bag of dims: Training-free standard-basis reading of transformer features, 2026. Manuscript.
- nostalgebraist. Interpreting GPT: The logit lens, 2020. LessWrong.
- Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In ICML, 2024. arXiv:2311.03658.
- Qwen Team. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2024.
- Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, et al. Open problems in mechanistic interpretability. arXiv preprint arXiv:2501.16496, 2025.
- Adly Templeton et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 sonnet. Transformer Circuits Thread, 2024.
- Qinan Yu, Jack Merullo, Ellie Pavlick, et al. Interpreting key mechanisms of factual recall in transformer-based language models. arXiv preprint arXiv:2403.19521, 2024.
- Zhang et al. Do all autoregressive transformers remember facts the same way? a cross-architecture analysis of recall mechanisms. arXiv preprint arXiv:2509.08778, 2025.
- Andy Zou et al. Representation engineering: A top-down approach to ai transparency. arXiv preprint arXiv:2310.01405, 2023.

A Per-model correctness monitor

Table 2 breaks down the correctness monitor of Section 8 by model. The monitor reads the late-band sign-commitment margin at a fixed penultimate layer (training-free) and predicts whether the model’s own greedy answer is correct. We report the training-free sign-margin AUC alongside a plain logit-lens confidence read on the identical states. Two patterns explain the 0.65–0.94 range in the main text: (i) the monitor AUC rises with the model’s own base accuracy (the weak Pythia models sit at the floor, the strong Mistral/Llama/Gemma models at the ceiling), and (ii) the sign margin and the logit-lens read are comparable, with the sign margin leading on 7 of 12 models and the logit lens on 4 (one tie) and mean AUCs within 0.01, so the sign monitor’s contribution is being training-free and fixed-layer rather than beating the lens. The open-domain transfer is weak for both reads: on a small web-questions probe the late-layer sign read scores AUC $\approx$ 0.56 and the logit-lens read $\approx$ 0.50, both near chance.

Model	Base accuracy	Sign-margin AUC	Logit-lens AUC
Pythia-410M	0.67	0.67	0.62
Pythia-1.4B	0.69	0.65	0.61
Pythia-2.8B	0.81	0.75	0.71
Pythia-6.9B	0.87	0.72	0.65
Qwen2.5-0.5B	0.76	0.66	0.66
Qwen2.5-1.5B	0.91	0.80	0.85
Qwen2.5-3B	0.94	0.75	0.87
Qwen2.5-7B	0.97	0.82	0.81
SmolLM2-1.7B	0.94	0.79	0.84
Mistral-7B	0.95	0.94	0.92
Llama-3.1-8B	0.97	0.90	0.86
Gemma-2-9B	0.97	0.93	0.95
Mean	0.87	0.78	0.77

Table 2: Correctness monitor by model (fixed penultimate-layer read). The training-free sign-margin AUC tracks the model’s base accuracy (floor 0.65 at Pythia-1.4B, ceiling 0.94 at Mistral-7B), and is comparable to a plain logit-lens confidence read on the same states (the sign margin leads on 7 of 12 models, the lens on 4).