WHERE DOES THE VERBALIZABLE WORKSPACE BAND BEGIN? AN AUDIT OF THE JACOBIAN-LENS INSTRUMENTS ACROSS OPEN-WEIGHT FAMILIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Claims about where verbalizable representations occupy a model's depth depend on an instrument that can locate the region and distinguish known capacities. We audit the Jacobian lens (J-lens), a map from intermediate states into vocabulary space, used by Gurnee et al. (2026) to place a mid-depth workspace band at roughly 38% of depth. We test five measurements, four band signatures and one occupancy estimator; one signature does not pass our validation conditions and one estimator depends on an unstated convention. One band signature, excess kurtosis, changes sign between two fits of one model differing only in fit corpus. Whether occupancy, under every operator we test, can separate 25 from 40 planted directions depends on an unembedding convention the original leaves unstated: it cannot on five of six open-model J-lens dictionaries uncentered, and can on all six centered. What survives is a J-lens effective-dimensionality transition, not a fixed workspace-band location: with reference-free averaged-Jacobian lenses and one rule fixed in advance, its onset ranges from 28.6% to 78.7% across seven families on 22 screened model–lens artifacts. Its location varies across model families and model–lens artifacts and depends on threshold and locator choice; open families begin both earlier and later than 38%, so comparison against that frontier-scale figure is uninformative.

1 INTRODUCTION

Gurnee et al. (2026) characterize transformer language models as organizing a small, privileged set of representations into a mid-depth "global workspace", a capacity-limited band of content that is verbalizable, reasoned-with, and broadcast to downstream computation, echoing the global-workspace theory of access in cognitive science (Baars, 1988; Dehaene & Naccache, 2001). They operationalize this characterization through an averaged-Jacobian "J-lens" readout and a set of band signatures measured across model depth, and report that the band runs from roughly a third of the way through the network to shortly before the output, with a capacity of about 25 simultaneously active directions.

The reported depth and capacity leave open whether the onset travels across open families under one rule. The readout matters too: the logit-lens proxy ($J = I$; nostalgebraist, 2020) is output-aligned by construction, so the onset must also survive the correctly-fitted averaged Jacobian. We take the original paper's own signatures and apply them with a reference-free fitted J-lens (Figure 1): five lenses we fit ourselves, on Gemma-2 (2B, 9B, 27B), Qwen2.5-7B and Qwen3-32B, placed alongside every published comparison lens we can read at the same resolution: 22 screened model–lens artifacts across seven families.

Under a fixed rule, the onset is not portable across open families: the original signatures yield a J-lens effective-dimensionality transition, not a located workspace band, with onset from 28.6 to 78.7% of depth across the 22 screened artifacts and seven families (19.7 points across our own five fits). And whether the capacity estimator is identified turns on a convention the original leaves unstated: uncentered, the procedure behind the published 25 cannot separate 25 planted directions

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.

from 40 on five of six open-model J-lens dictionaries; centered, it separates them on all six. This bounds the estimator, not the frontier-scale model.

Establishing this required instruments we built and validated: a reference-free fitted J-lens, checked by finite differences, a last-layer identity invariant and a per-model positive control; a locating rule fixed in advance, whose onset is unmoved in all 400 replicates when the lens's fitting prompts are resampled; a within-model decomposition placing the depth profile in the Jacobian, not the unembedding; a continuous locator recovering known transitions to a median error of 0.479 depth points; and a planted-count gate sealed in advance. Three of the four signatures organize into the original sensory, band and motor regions. Of the five measurements, one signature fails validation and one estimator is convention-dependent, both as properties of the estimators rather than of the models. Excess kurtosis changes sign between two fits of one model; on five Gemma-2-2B fits only the largest trim tested makes it reproducible, clearing the bar by 0.019. Throughout we separate variation across artifacts at a fixed rule (Section 4) from variation across rules at a fixed model (Section 4.1).

2 Related Work

Jacobian-lens interpretability. The averaged Jacobian is a first-order, layerwise readout of the residual stream, and Gurnee et al. (2026) use it, with a family of band signatures, to characterize a mid-depth verbalizable-workspace band on a closed, frontier-scale model. Open-weight behavior under that lens is no longer uncharacterized: Wang & Reid (2026) extend it to looped and depth-recurrent architectures, with a standard-layer model as baseline, and report that a workspace forms in the iterated part of each while recurrence changes how it can be accessed. What remains uncharacterized is whether the depth at which the band *begins* is portable across families under a rule fixed in advance, and we ask it by applying the original account's own signatures to the population of Section 1, with a reference-free fitted J-lens in place of the logit-lens proxy. Where our numbers differ, the difference is one of location and of instrument identifiability, at a scale the original work does not claim to describe. Attending to the instrument is not incidental: interpretability findings can depend on measurement and setup choices (Bricken et al., 2024).

Readout alignment and the logit lens. The logit lens (nostalgebraist, 2020) projects each layer directly through the unembedding; it sets $J = I$ and is output-aligned by construction, so it cannot resolve mid-depth structure departing from the output direction. Its brittleness motivated learned successors: the tuned lens (Belrose et al., 2023), an affine probe fitted per layer to predict the output distribution rather than estimated as the model's own Jacobian; Future Lens (Pal et al., 2023), which reads several tokens ahead from one hidden state; and Patchscopes (Ghandeharioun et al., 2024), which subsumes vocabulary-space projections into a single patching framework. All are fitted or prompted decoders. The averaged Jacobian is none of these: it is estimated as the model's own derivative, which is why it admits the reference-free gates of Appendix B and why $J = I$ is a control we can run rather than a baseline we must trust.

Sparse dictionaries over the residual stream. A J-lens readout decomposes activations against a dictionary of vocabulary-sized directions, which places it alongside a line of work that learns overcomplete dictionaries for the residual stream directly: sparse autoencoders trained to recover monosemantic features (Bricken et al., 2023; Huben et al., 2024), scaled to frontier models (Templeton et al., 2024), and extended by transcoders that replace a dense sublayer with a wider sparse one (Dunefsky et al., 2024). Those dictionaries are learned; ours is induced by the unembedding and never trained. The identifiability question applies to both: any capacity read off a highly coherent overcomplete dictionary inherits the recovery limits of Appendix G. The theory that would state the condition before such a count is reported has only recently appeared (Ivanitskiy et al., 2026).

Depth-resolved analyses, and dimensionality across depth. Measuring where in depth a property emerges has a substantial history: edge probing recovers a pipeline-like ordering of linguistic tasks in BERT (Tenney et al., 2019; Rogers et al., 2020), mechanistic work localizes circuits such as induction heads to characteristic depths (Olsson et al., 2022), and Lad et al. (2025) argue from layer deletion and swapping for four stages of inference, a partition close in spirit to the three-region one we score against. Closest to signature (d) is the intrinsic-dimension literature: Ansuini et al. (2019)

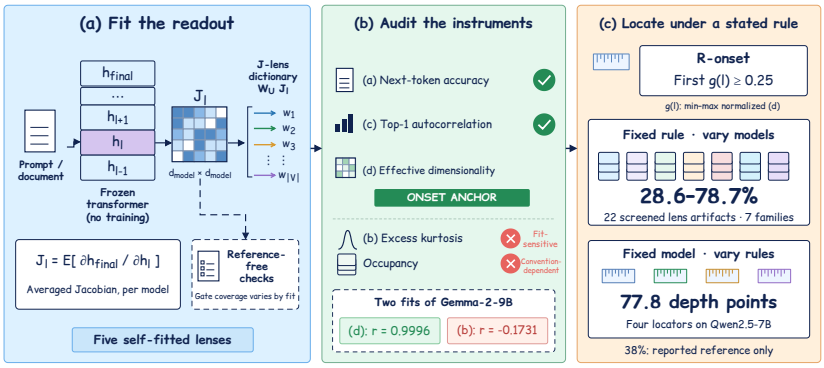

Figure 1: What this paper does: fit a lens per model, audit the five measurements read off it, then locate the transition under one rule fixed in advance. Schematic; measured values appear in Sections 3–4. "Convention-dependent": the count is not identified on five of the six dictionaries gated with the unembedding uncentered, and is identified on all six with it centered (Section 5.2).

find the ID of deep-network representations rising then falling with depth, and Valeriani et al. (2023) an expand–contract ID profile across depth in large transformers trained on protein and image tasks. A depth-varying dimensionality is therefore not our finding: (d) is a spectral statistic of the *dictionary* $W_U^\top J_\ell$, not of activations, our result is the portability of a threshold crossing on it, and whether that crossing coincides with the ID structure, or is distinct from it, is untested here.

3 Methodology

3.1 Averaged-Jacobian lens

Portability cannot be tested until the lens and locating rule are fixed. Figure 1 summarizes the lens pipeline. For a fitting prompt x, let P_x be its valid token positions. At layer l we estimate the full token Jacobian by reverse-mode backpropagation through the attention blocks (Vaswani et al., 2017), seeding each output coordinate at all valid target positions. The per-prompt matrix averages over source positions and sums their causally reachable targets; the fitted lens is

$$J_l = \mathbb{E}_x \left[\frac{\sum_{p \in P_x} \sum_{p' \in P_x : p' \geq p} \partial h_{\text{final},p'} / \partial h_{l,p}}{|P_x|} \right],$$

where the expectation is over fitting prompts. For a row-vector residual state h_l, the J-lens readout transports first and then applies the model's final normalization and unembedding: $\text{ln_final}(h_l J_l^\top) W_U$ (followed by the model's logit soft-cap where present). For decompositions of model activations we apply that final normalization per token and subtract the per-layer mean over included tokens, removing the common cross-token direction.

The fitting regime is $n = 464$ prompts for Gemma-2-9B (Gemma Team, 2024) and Qwen2.5-7B (Qwen Team, 2024), and $n = 150$ for Gemma-2-2B; Table 1 gives all five with their validation coverage. On Gemma-2-2B, the fit with a retained convergence curve, the onset is converged in fit size and (d)'s residual sensitivity is bounded at one layer (Appendix J).

Three reference-free gates establish lens correctness: fp64 central-difference agreement on GPT-2 (Radford et al., 2019) to relative error 1.26×10^{-9}–5.25×10^{-9}, the fitted Jacobian approaching identity toward the target layer, and signature (a) rising toward the output (Appendix B).

3.2 Band signatures

A correct lens still leaves the measurements read through it to validate. Following the original paper, we compute four signatures across depth, reindexed to $[0, 100]\%$: (a) top-1 next-token accuracy of the lens readout (*actual-next-token* convention throughout, i.e. agreement with the ground-truth next token; we do not mix this with the model-argmax convention used by the positive-control gate,

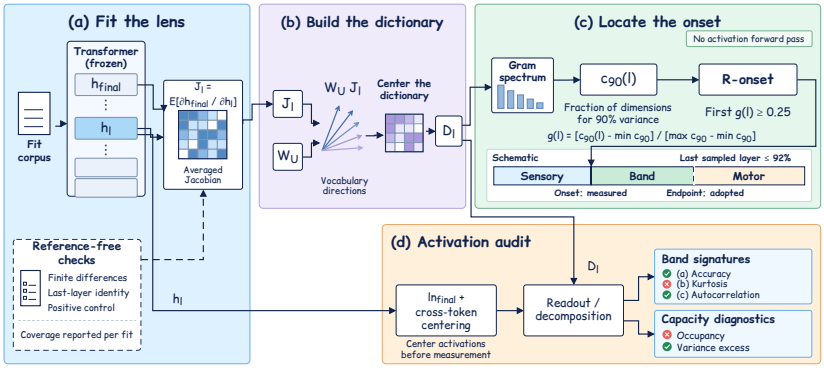

Figure 2: From a fitted J-lens to a rule-defined onset. $W_U J_l$ is centered over vocabulary rows to give D_l; $c_{90}(l)$ is read from its Gram spectrum, min–max normalized to $g(l)$, and R-onset returns the first depth with $g(l) \geq 0.25$. The upper edge is adopted, not measured. Panel (d) applies to model activations; the planted-count gate decomposes raw synthetic ones (Appendix G.4).

Appendix B); (b) excess kurtosis of the lens readouts; (c) top-1 autocorrelation across sequence positions; and (d) the spectral fraction of residual-stream dimensions needed to capture the prescribed variance share across the J-lens dictionary. Specifically, with $D_l = W_U^\top J_l$, we subtract its mean vocabulary row to obtain $\widetilde{D}_l$. We call that subtraction *U-centering*, and the unsubtracted form *U-raw*; it is distinct from centering the activations, which Appendix G.4 treats separately. Signature (d) is computed U-centered throughout; occupancy and variance-excess are reported U-raw unless marked, and Section 5.2 reports both. If $\lambda_{l,i}$ are the eigenvalues of $\widetilde{D}_l^\top \widetilde{D}_l$ in descending order, then

$$c_{90}(l) = \min\left\{ k : \sum_{i \leq k} \lambda_{l,i} \geq 90\% \sum_i \lambda_{l,i} \right\} / d.$$

Signature (d) needs no forward passes (only the unembedding and the lens), and we compute it through the exact Gram identity so that the $(\text{vocab} \times d)$ dictionary is never materialized. It is also, by a wide margin, the most stable of the four across independent fits: between two of our own Gemma-2-9B fits that differ only in fit corpus, $r = 0.9996$ for (d) against $r = -0.1731$ for (b) (Section 5.1). That stability is why we use it, and only it, to locate the effective-dimensionality transition.

3.3 Locating the transition: a rule fixed in advance

A transition claim is only as good as the rule that locates it, so we state ours before applying it and report the result whether or not the signatures then agree.

R-onset. Signature (d) is the anchor, which makes any test scored against the resulting operational region partly circular; Section 4.1 therefore repeats the exercise with locators that never touch (d). Min–max normalize signature (d) over the measured depths, $g(l) = \big(c_{90}(l) - \min_l c_{90}(l)\big) / \big(\max_l c_{90}(l) - \min_l c_{90}(l)\big)$. The onset is the first layer with $g(l) \geq 0.25$: the depth at which effective dimensionality first exceeds a quarter of its total range.

R-premotor. As sealed in advance, the pre-motor boundary was the layer before the largest single-step increase in signature (d). That rule failed: it is unstable to fit corpus and returns single-layer bands on two models (Appendix D), too much play to carry a claim about where the band ends. We make none, and adopt instead the endpoint the original paper itself quotes, 92% of depth, taken from the target paper rather than derived. After the failure the onset is the only band edge our data chooses, and wherever signatures are scored against a three-region partition that partition's upper edge is the audited account's, not ours.

For scoring only, we retain the original account's region names: SENSORY below the onset, BAND from the onset to the pre-motor boundary inclusive, and MOTOR above it, indexed by layer so

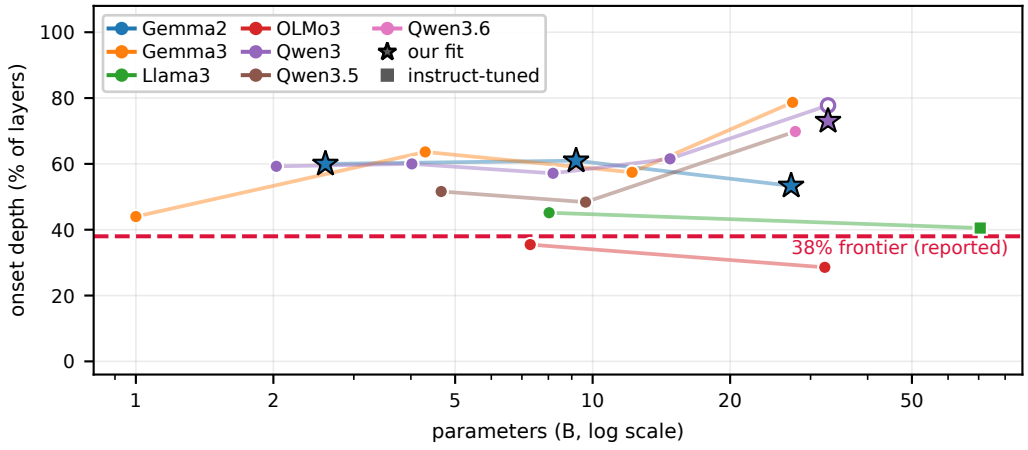

Figure 3: The onset under R-onset across the 22 model–lens artifacts that survive the resolution screen: seven families, one rule, one threshold. Stars are lenses we fitted ourselves (four of five; Qwen2.5-7B is not on the ladder), filled circles published comparison lenses, the open circle the 80-prompt mid-fit Qwen3-32B checkpoint, squares instruction-tuned artifacts. The dashed line is the 38% reported at frontier scale, a reference rather than a target: artifacts from different families fall on both sides. The vertical extent, not the position of any one artifact, is the result.

artifacts on different depth grids are cut identically. These are operational cuts, not established workspace locations; Appendix A collects this paper's terms of art in one place. We fix in advance what it means for each signature to match the predicted shape: (a) attains its maximum in the motor region with band above sensory; (b) band above sensory; (c) attains its maximum inside the band, above both neighbours; (d) band above sensory and motor above band; occupancy elevated inside the band relative to both neighbours; variance-excess never exceeding 10%.

The rule needs a fitted Jacobian. Under the logit-lens proxy it has nothing to locate: signature (d) reads the spectrum of $J_\ell^\top M J_\ell$, where the centered unembedding Gram M carries no ℓ, so $J_\ell = I$ makes the profile exactly constant and R-onset undefined. Run as a control on all three lenses with a stored Jacobian, that is what it returns: zero spread in frac90 across depth to machine representation, R-onset NaN (Appendix I). The depth profile is carried by the Jacobian, not the unembedding.

The common fitting regime and its tokenizer consequence are recorded in Appendix B; prompt counts and validation evidence differ across the five fits, as Table 1 makes explicit.

4 Where the transition begins

Section 3 fixed one rule, one estimator and one regime, enough to locate the transition in a single model. Whether that number describes anything beyond that model is the question this section asks, and the rule does not change to answer it: R-onset stays the first depth at which min–max-normalized signature (d) exceeds 0.25 of its range, fixed before any outcome was inspected. What changes is the population. Table 1 records what each fitted lens carries and how each was validated.

Our own five fits disagree by 19.7 points. Applying R-onset to the five lenses of Table 1 gives 53.3% (Gemma-2-27B), 60.0% (Gemma-2-2B), 61.0% (Gemma-2-9B), 66.7% (Qwen2.5-7B) and 73.0% (Qwen3-32B). The estimator, corpus source, sequence length, depth sampling, rule, and threshold are shared, but prompt count and validation evidence are not; the two largest fits carry no gate artifact. The resulting 19.7-point spread is descriptive of this self-fitted subset, not a matched-model control; nor is it fit-sampling noise, since resampling a lens's prompt shards returns the same onset in all 400 replicates on both lenses where shards were retained (Appendix J). Read on the first three alone the spread is 6.7 points; adding the two further artifacts nearly triples it, so the narrow figure was a property of that three-artifact sample rather than a constant of the architecture.

Across families the onset spans 50.1 points. Signature (d) needs no forward passes, so an onset can be read at the same rule and resolution wherever a lens has been published. Pooling four of our five fits with the published comparison lenses gives 26 model–lens artifacts; 22 survive the resolution screen below, across seven families. Qwen2.5-7B carries the signature and capacity work but is not on the ladder as assembled; its onset of 66.7% falls inside the reported range, and its inclusion would not change it. Their onsets run from 28.6% (OLMo-3-32B) to 78.7% (Gemma-3-27B), a spread of 50.1 points (Figure 3). Every artifact carries the onset alone: none is re-cut for signatures and none enters a match count (Appendix C). The range does not rest on its least certain endpoint: without Gemma-3-27B (mirrored weights, a 37.7-point sweep, the largest locator standard error, 24.98) it is 28.6–77.8%, without the 80-prompt Qwen3-32B checkpoint too 28.6–73.0%, and no single removal leaves under 43.2 points.

The resolution screen. R-onset is a threshold crossing on a curve sampled at discrete depths, so its precision is bounded by the grid. The screen is on nominal spacing, 25 depths over the layer count; samples land on integer layers, so consecutive steps vary about that value. We admit an artifact only if its nominal grid is at least as fine as 4.2 points per step, applied as a criterion rather than case by case. Four of the 26 fail and are excluded from every range above; Appendix C gives the cases. Excluding the two low-fit published artifacts leaves 20 and the range unchanged.

Within-family movement, and a matched-scale control. Within-family spans run 4.7 points (Llama-3), 6.9 (OLMo-3), 7.6 (Gemma-2), 20.6 (Qwen3), 21.5 (Qwen3.5) and 34.7 (Gemma-3). Gemma-3's swing is not a scale effect: its 1B-to-4B step changes vocabulary and unembedding path too (Appendix C). At parameter counts matched to within a factor of 1.5, six models from six families between 27 and 33B span 28.6–78.7%, the entire seven-family range at one parameter count; five more between 7 and 10B, from five families, span 25.5 points. Across the 22 screened model–lens artifacts the correlation of onset with $\log_{10}$ parameters is $r = +0.16$, descriptive on n=22 with three duplicated parameter counts. Six artifacts is few.

The spread against artifact quality. The sharpest objection to the range is that it is an artifact-quality effect: the three fits passing all three gates span only 6.7 points, and the 50.1-point range appears once artifacts we cannot validate to that standard are added. Two comparisons bear on it. Where one model is read through two independently produced lenses, its onset barely moves: Gemma-2-27B gives 53.3% under our own n=464 fit and 53.3% under an independently published n=780 lens, Gemma-2-9B 61.0% against a comparison lens's 65.9%, and Qwen3-32B 73.0% against an 80-prompt checkpoint's 77.8%. And the spread is not confined to the validated–unvalidated boundary: the four Gemma-3 artifacts are published lenses read through the same mirrored weights, indistinguishable in validation, and span 34.7 points. Where we can measure it, lens provenance moves the onset by points, not tens.

The gated subset is three models from two families, none above 9.2B, and every artifact outside it is either a published lens or one of our two largest fits, so validation quality, family and scale are not separable here: the 6.7-point figure cannot be told apart from "three closely related small models", and we do not claim a larger gated subset would have stayed narrow.

4.1 Locators that do not use signature (d)

Section 4 held the rule still and varied the models; this section does the reverse. Under three further locators fixed in advance, none touching (d), the four rules disagree *within* one model by 32.0, 31.7 and 77.8 points on Gemma-2-2B, Gemma-2-9B and Qwen2.5-7B, of the same order as the first axis, where one rule across seven families spans 50.1 points. The two spreads compound rather than explain one another, and neither licenses a statement about where the transition begins in general. Two of the three independent locators agree with (d) in direction and disagree in magnitude, so on these three models the onset is not an artifact of anchoring on (d); the third fails, placing an onset in the sensory region. Appendix E gives the three rules, their per-model onsets and that failure (Table 4).

4.2 What the frontier-scale figure can and cannot be compared with

The depth axis transfers: the original work reports on 25 evenly spaced layers reindexed to $[0, 100]$, the convention we adopt. Two things defeat the comparison anyway, both consequences of Section 4 rather than of any re-analysis of the frontier figure. The population brackets the figure: OLMo-3 begins below 38% on both artifacts we can read, Llama-3 and Gemma-3-1B between 40.5 and 45.2%, and every other Gemma and Qwen artifact at 48.4% or later, so no single figure from a single model is early or late against it, and averaging over families would only hide which side each falls on. And the procedure was never like-for-like: the 38% figure is a qualitative reading over several analyses where ours is one explicit rule on one signature, and their per-layer values are not published, so we cannot re-derive theirs under R-onset. The gap between a threshold crossing and a qualitative reading is of the same order as any difference such a comparison would report: four rules on one model span 77.8 points (Table 4). We withdraw the comparison in either direction.

5 Whether the signatures agree on that band

Section 4 used one signature and rule to locate the transition and define a candidate region. Whether it matches the original band claim depends on the other signatures; of the five measurements, one signature does not pass our validation conditions and one estimator depends on an unstated convention. Signature (b) is not reproducible in sign across two fits of one model, and whether the tested occupancy estimators distinguish planted counts depends on the unembedding convention. We audit those two first, then report what remains.

5.1 Signature (b) fails as a locator under the tested fits

Signature (b) anti-correlates at $r = -0.1731$ between two of our own Gemma-2-9B fits differing only in fit corpus, where (a), (c) and (d) agree above 0.98, and ten vocabulary entries of $256{,}000$ decide which side of zero its band-minus-sensory contrast falls: it fails by disagreeing with itself. It matches on 3 of 6 profiles, which is what a measurement looks like when a handful of extreme values dominate it, so we read neither a band nor its absence off signature (b) on Gemma-2 or Qwen2.5 up to 9B. Appendix F gives both diagnostics, the structural reason, and the trimming curve.

5.2 Occupancy's identifiability depends on an unstated convention

Occupancy, the original account's capacity measure, counts simultaneously active J-lens directions, about 25 inside the band. Its dictionary is the rows of $W_U J_\ell$, and whether W_U is centered, its vocabulary mean subtracted, decides whether the count is identified. The original account leaves the convention unstated: its paper does not specify it and its released code contains no occupancy estimator. We run the planted-count gate (Appendix G.5) under both. *Uncentered*, it fails on five of six open-model J-lens dictionaries at 48.0–90.2% depth, where no operator, the original account's gradient pursuit included, separates 25 planted directions from 40; the exception is Gemma-2-9B at 90.2%. Under the uncentered construction, identifiability varies with depth and family, without a threshold common to the families we test: Gemma-2-2B separates from somewhere in $(80.0, 88.0]\%$, Gemma-2-9B from somewhere in $(61.0, 85.4]\%$, and Qwen2.5-7B at no depth we test up to 92.6% (Appendix G.6). *Centered*, every dictionary tested separates, all six (Table 7), so the pattern is the uncentered construction's. The models' own activations split the same way: uncentered, median occupancy never leaves single digits against a reference capacity of 25; centered, it reaches 10 inside the Gemma-2-2B band and 20–25 inside Qwen2.5-7B's (Appendix G.3). Occupancy's identifiability on these dictionaries is therefore a property of the unembedding convention, and a count read off it, the published 25 included, may reflect that convention as much as the model. That bounds the estimator on our open-model dictionaries and on synthetic sparse structure, the case most favourable to it; it is not a measurement of the frontier-scale model and neither refutes nor confirms the number there.

5.3 Where the surviving signatures place the region

With those two set aside, what remains is where the surviving signatures place the region, and how much of that agreement is earned rather than built in.

Table 2 re-cuts every signature against the per-model bands of Section 4; Figure 5 plots the profiles. Signatures (a), (c) and (d) match the predicted shape on all three of our own fits. Across all six model–lens profiles we hold (three primary fits and three comparison lenses): (a), top-1 next-token accuracy, matches on 6 of 6; (c), top-1 autocorrelation, on 5 of 6 (the exception is a comparison lens whose (c) peak falls one layer outside its own band, a near miss a four-point shift in the onset would recover; Appendix E); (d), effective dimensionality, on 6 of 6; and (b), excess kurtosis, on 3 of 6.

Signature (d) matches on all 6 profiles, but half of that test is circular and we separate the halves. "Band above sensory" is close to guaranteed for the curve the onset was read off; the informative half is "motor above band", independent of the onset rule, and it holds on all 6 (Table 2).

5.4 Variance-excess, and what it does at 27B

Section 5.2 owed a replacement, and it is not a capacity estimate. Variance-excess (the reconstruction variance captured by the top-K directions in excess of a matched random baseline, at the median occupancy) does not count active directions: under the same gate it responds monotonically to the planted count but with *negative* ρ, falling as more are planted, and never saturates on the tested grid (Appendix G.5). We report it as a non-saturating diagnostic of support size, in occupancy's place but not in occupancy's role. In the band it peaks at 5.51% (at 84.6% depth) on Gemma-2-2B, 6.64% (at 88.1%) on Gemma-2-9B (a comparison lens; Table 5) and 1.65% (at 85.7%) on Qwen2.5-7B, and no depth on any of the three exceeds the 10% ceiling the original account states. All three use one unembedding convention (Appendix G.3).

The 27B artifact narrows the claim rather than extending it: the signal holds there but weakly, with full-depth maxima of 1.21% under our $n{=}464$ fit and 1.14% under the published $n{=}780$ lens, against 5.51% at 2B and 7.25% at 9B, the latter above the band and on the comparison lens (Figure 6). The series is not monotone and its three scales are not sampled comparably, so it is a statement about the 27B artifact rather than a trend in scale.

5.5 What the localization survives

One signature failed and three held, and occupancy, the capacity measure, is identified under one unembedding convention and not the other. Whether the onset the survivors produce is an artifact of how we fit them is the last question about the instrument, and it is not: the localization survives every confound we tested on Gemma-2 and Qwen2.5 up to 9B: fit size (with a matched-n control at 9B), activation convention, DC contamination, the decomposition budget k, and a $2{\times}2$ wikitext $\times$ GSM8K fit-by-eval design. The onset is stable across fit corpora: 61.0% on Gemma-2-9B for both our wikitext and our reasoning fit. Appendix J gives every number, including one residual GSM8K-fit $\times$ GSM8K-eval interaction reported as unresolved and bounded.

6 Discussion

The surviving signatures support an effective-dimensionality transition rather than a constant workspace-band location, though they reproduce the original account's three-region organization. Artifacts at one rule and rules at one model produce spreads of the same order. Under our validation conditions, (b) cannot locate the transition, and whether occupancy's estimator can tell the published 25 from 40 depends on an unstated convention, failing on five of six open-model dictionaries uncentered and on none centered: a limit on the estimator, not the frontier-scale model.

What the surviving instrument is for. The audit is not only negative. Signature (d) is the most reproducible measure we have: two independent fits of one model agree at $r = 0.9996$, and its onset on Gemma-2-2B is fixed from $n{=}20$ prompts on. It supports a within-model, within-rule statement, and not a portable number: two users fixing different thresholds, or reading different families, obtain onsets differing by more than most effects such numbers argue for.

The model–lens confound. The onset varies across artifacts, not isolatably by family, because family co-varies with layer count, width, attention architecture, tokenizer size, embedding tying, training and lens fitting; Gemma-3's 34.7-point span, which crosses the architecture change between

its 1B and 4B artifacts, is the same confound. Separating these needs one factor varied and the rest held, which no lens population provides. Appendix I does that inside single models, finding the depth profile carried by the Jacobian rather than the unembedding on all three lenses with a stored Jacobian; it does not reach the cross-family question.

Where the J-space component becomes sufficient. The transition is a spectral statistic, so whether it marks anything the model does is a separate question. One probe, sealed in advance, tests it on Gemma-2-2B: swapping the J-space component of the final-position state from one capital-city prompt into another, over 96 pairs, transfers the source's answer from 80.0% depth, while random-dictionary and complementary swaps transfer nothing at any depth (Appendix L). The component is therefore sufficient to carry the answer, and specifically so; but it becomes sufficient 20 points after the (d) onset, and part of that gap belongs to the site, where even the whole state transfers only from 72.0%. One model, one site, one relation: the onset is not where the component becomes sufficient, and a general functional account remains untested.

What follows for a reported onset. A single number needs its rule and population, and earlier-or-later claims require the same rule and comparable depth resolution. The signature that fails our validation conditions and the estimator that depends on an unstated convention are both facts about instruments, not claims about the models.

7 Scope and Limitations

The screened ladder spans 1.0–70.55B; the frontier scale has no open weights. Validation is uneven: the two largest of our five fits carry no gate artifact, and eight of the 26 artifacts, seven of them screened in, use ungated full-precision mirrors (Table 3). That gap cannot be closed after the fact: the per-layer Jacobians of those two fits were not retained, so the identity gate cannot be run on them at all. The 22 artifacts are screened on depth resolution only, with fit size (80 to 1000 prompts), tuning and the confounds of Section 6 left free, so the ladder is descriptive, not a matched control.

Two further limits bound the claim rather than the population. Only the onset comes from our data; the pre-motor endpoint is taken from the original paper after our rule proved unstable (Section 3.3). The onset is read off signature (d), so only (d)'s "motor above band" half is independent, and no onset we report is rule-independent (Section 4.1).

Of the confounds we could test, fit size is discharged for Gemma-2-2B at the threshold we fixed in advance but not at every threshold in the sweep (Section 3). The fitting corpus is wikitext-103 plus a reasoning re-fit, not an exhaustive distributional sweep. Three things remain open. The Gemma-2-9B capacity artifact uses a comparison lens, recomputation on our own fit having been blocked by storage, a further gap in the reference-free chain beyond the two ungated fits. Why the real dictionaries fail harder than coherence-matched synthetics is unresolved (Appendix H). And the GSM8K-fit$\times$GSM8K-eval residual remains unresolved. Finally, the functional probe of Section 6 is one model, one site and one relation, and a steering pilot failed its gate.

8 Conclusion

Under reference-free J-lenses and one fixed rule, an effective-dimensionality transition appears in all 22 screened artifacts in seven families, with onset at 28.6–78.7% of depth, 19.7 points of that across our five lenses, and as much across rules on one model. Open families straddle the frontier-scale 38%, so we make no comparison. Signatures (a), (c) and (d) cohere around the operational region; (b) does not locate the transition. Whether the capacity estimator separates the published 25 from 40 depends on an unembedding convention the original leaves unstated: it fails on five of six open-model dictionaries uncentered and on none centered, a bound on the estimator, not the frontier-scale model. An onset requires its rule and population.

AI use statement

We used an AI assistant to aid and polish the writing of this paper. Every claim in the text, tables and figures is the authors' responsibility.

No experimental result was produced by a generative model; every number is computed from model weights by the code described in the appendices.

Reproducibility Statement

Every model we read is open-weight and named with its Hugging Face identifier (Table 1, Appendix C). The lens estimator is given in full in Section 3, the three reference-free validation gates in Appendix B, the locating rule and its threshold in Section 3.3, and the planted-count gate with its degrees of freedom in Appendix G. A lens can be refitted from that specification; we distribute no fitted lens and no model weights, so the artifacts behind the reported numbers are ours rather than the reader's, and the numbers should be read accordingly.

Where we say a plan, rule or bar was sealed or fixed in advance, we mean an analysis plan written and committed internally before the run it judges. None was publicly registered; a citable public registration is future work.

References

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1905.12784.

Bernard J. Baars. *A Cognitive Theory of Consciousness*. Cambridge University Press, Cambridge, 1988.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023. URL https://arxiv.org/abs/2303.08112.

Thomas Blumensath and Mike E. Davies. Gradient pursuits. *IEEE Transactions on Signal Processing*, 56(6):2370–2382, 2008. doi: 10.1109/TSP.2008.916124.

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, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, October 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Trenton Bricken, Jonathan Marcus, Siddharth Mishra-Sharma, Meg Tong, Ethan Perez, Mrinank Sharma, Kelley Rivoire, and Thomas Henighan. Using dictionary learning features as classifiers. Transformer Circuits Thread, October 2024. URL https://transformer-circuits.pub/2024/features-as-classifiers/index.html.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Stanislas Dehaene and Lionel Naccache. Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, 79(1–2):1–37, 2001. doi: 10.1016/S0010-0277(00)00123-2.

David L. Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003. doi: 10.1073/pnas.0437847100.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable LLM feature circuits. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2406.11944.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. URL https://arxiv.org/abs/2408.00118.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Proceedings*

of the 41st International Conference on Machine Learning (ICML), 2024. URL https://arxiv.org/abs/2401.06102.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, T. Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. Transformer Circuits Thread, July 2026. URL https://transformer-circuits.pub/2026/workspace/index.html. arXiv:2607.15495.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Michael I. Ivanitskiy, John Jasper, Emily J. King, and Dustin G. Mixon. Towards a mathematical theory of superposition. *arXiv preprint arXiv:2608.27540*, 2026. URL https://arxiv.org/abs/2608.27540.

Vedang Lad, Jin Hwa Lee, Wes Gurnee, and Max Tegmark. Remarkable robustness of LLMs: Stages of inference? In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2406.19384.

Charles L. Lawson and Richard J. Hanson. *Solving Least Squares Problems*. Prentice-Hall, Englewood Cliffs, NJ, 1974.

Stéphane G. Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993. doi: 10.1109/78.258082.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1609.07843.

Neel Nanda and Joseph Bloom. TransformerLens. https://github.com/TransformerLensOrg/TransformerLens, 2022.

nostalgebraist. Interpreting GPT: The logit lens. LessWrong, August 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022. URL https://arxiv.org/abs/2209.11895.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025. URL https://arxiv.org/abs/2508.10925.

Koyena Pal, Jiuding Sun, Andrew Yuan, Byron C. Wallace, and David Bau. Future Lens: Anticipating subsequent tokens from a single hidden state. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, 2023. URL https://arxiv.org/abs/2311.04897.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Qwen Team. Qwen3.5: Model card. Hugging Face, 2026a. URL https://huggingface.co/Qwen/Qwen3.5-27B.

Qwen Team. Qwen3.6-27B: Model card. Hugging Face, 2026b. URL https://huggingface.co/Qwen/Qwen3.6-27B.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI technical report, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. doi: 10.1162/tacl_a_00349.

Team Olmo. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025. URL https://arxiv.org/abs/2512.13961.

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, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel 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. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, 2019. doi: 10.18653/v1/P19-1452.

Joel A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50:2231–2242, 2004. doi: 10.1109/TIT.2004.834793.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 51234–51252, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a0e66093d7168b40246af1cddc025daa-Abstract.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wenlong Wang and Fergal Reid. Looped transformers under the Jacobian lens: Does the global workspace survive recurrence? *arXiv preprint arXiv:2609.01924*, 2026. URL https://arxiv.org/abs/2609.01924.

APPENDIX

A TERMS OF ART USED IN THIS PAPER

Several terms below are ours rather than standard, and two are easy to confuse with each other. They are defined once here and used only in these senses.

Table 1: The five lenses we fitted ourselves. All share one estimator and the frozen matched regime (wikitext-103, SEQ=128, SKIP=16, 25 source layers); only Gemma-2-2B departs, at n=150. The three above the rule pass all three validation gates (Appendix B). The two below it have no gate artifact at all, resting instead on convergence plus agreement with a published lens on signature (d), the most lens-stable signature we measure and so the weakest evidence of correctness. mirror marks weights read through an ungated full-precision mirror, never compared tensor-by-tensor against the gated original (Appendices B and C).

Model	fit n	validated	weights	carries
Gemma-2-2B	150	three gates	`google/gemma-2-2b`	onset, signatures, capacity
Gemma-2-9B	464	three gates	`google/gemma-2-9b`	onset, signatures
Qwen2.5-7B	464	three gates	`Qwen/Qwen2.5-7B`	onset, signatures, capacity
Gemma-2-27B	464	convergence + (d)	mirror	onset, variance-excess
Qwen3-32B	464	convergence + (d)	`Qwen/Qwen3-32B`	onset

term	sense used here
model–lens artifact	One (model, fitted lens) pair. Two artifacts can be the same model read through different lenses, so a spread across artifacts is not a spread across models.
rung	One model–lens artifact on the cross-family ladder, carrying an onset and nothing else: no rung is re-cut for signatures and none enters a match count.
ladder	The pooled set of rungs (Table 3): 26 before the resolution screen, 22 after. Descriptive, not a matched control.
R-onset	The locating rule fixed in advance (Section 3.3): min–max normalize signature (d) over the measured depths and take the first depth at which it reaches 0.25 of its range.
(d)-x25	R-onset named by the signature it reads, used in Table 4 to distinguish it from the three locators that never touch (d).
(a)-x25, (c)-x25	The same 0.25 crossing rule applied to signature (a) or (c) instead of (d).
(c)-peak	The depth of signature (c)'s inverted-U maximum.
SENSORY, BAND, MOTOR	The three regions signatures are scored against: below the onset, from the onset to the adopted 92% endpoint inclusive, and above it. Operational cuts named after the audited account's regions, not established workspace locations.
sweep	The width of a rung's onset excursion across the threshold grid $\{0.10, \ldots, 0.50\}$, reported for every rung. A wide sweep means a weakly determined onset.
screen	The resolution criterion admitting a rung only if its nominal depth grid is at least as fine as 4.2 points per step.
sealed in advance	An analysis plan, rule or bar written and committed internally before the run it judges. Not a public registration.
gate	A pass/fail check on an instrument, with its bar fixed before the run: the three lens-validation gates of Appendix B, and the planted-count gate of Appendix G.5.

B VALIDATION-GATE DETAILS

Common fitting regime. The rule alone does not make models comparable, so the fitting regime is held still as well. All fits use the same wikitext-103-raw-v1 source (Merity et al., 2017) (train split, native order, character-length filter followed by a 2000-character cut), $n = 464$, a maximum sequence length of 128 tokens with the first 16 positions skipped, each model's native tokenizer, and 25 evenly spaced source layers over $[0, \text{target})$ with target $= n_{\text{layers}} - 1$ (configuration frozen). We match *sequence length* (the model-intrinsic unit over which the Jacobian is computed), rather than word count; the resulting word-count difference across tokenizers is 2.6% at 128 tokens (Gemma 100.3 vs. Qwen 97.7 words/prompt), a small and deliberate consequence. Because roughly 35% of wikitext-103 paragraphs are shorter than the cap, 128 is a maximum rather than a uniform length.

Each of the three gates summarized in Section 3 is given below with its controls, its metric-scale caveats and its per-layer values.

(0) The two lenses with no gate. Gemma-2-27B and Qwen3-32B carry no gate artifact of any kind (Table 1). Their evidence is per-shard convergence ($\|J_l - I\|_F/\sqrt{d}$ at the top source layer

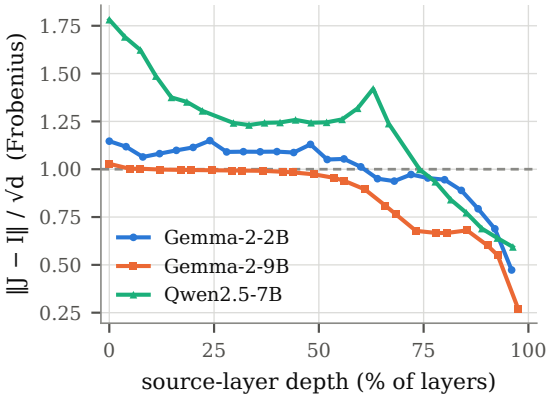

Figure 4: Last-layer identity invariant on our own fits. $\|J_l - I\|_F/\sqrt{d}$ decreases toward the target layer in every model, as the residual-stream structure requires. The dashed line marks the metric scale: the zero matrix scores exactly 1.0, so the invariant under test is the *decrease* toward the target rather than the absolute level.

is 0.6303/0.6304 on Gemma-2-27B and 0.4849/0.4834 on Qwen3-32B, two shards at n=232 each) plus agreement with an independently published lens on signature (d): $r = 0.9996$, max $|\Delta| = 0.0189$ on Gemma-2-27B and $r = 0.9992$, max $|\Delta| = 0.0205$ on Qwen3-32B. Since (d) is the most lens-stable signature we measure, that agreement is the weakest available evidence of correctness, and neither it nor shard convergence substitutes for the three gates below.

(1) Finite difference. GPT-2 is used here because it is the one model small enough to run fp64 finite differences at full precision: the choice is driven by numerical precision, not by any contrast with the models we localize. On GPT-2 at fp64 the averaged-Jacobian estimator matches central-difference numerical gradients to relative error 1.26×10^{-9} (L3), 1.98×10^{-9} (L6) and 5.25×10^{-9} (L8) at $\epsilon = 10^{-3}$; the src=tgt plumbing error is exactly 0.0. The worst case across the full ϵ grid is 9.07×10^{-8}, at the ends of the grid where the central difference is either truncation- or roundoff-limited; the values quoted in the main text are at $\epsilon = 10^{-3}$. Finite-difference on the 7B/9B lenses at bf16 is roundoff-dominated, so it is not a pass/fail test: the clean derivative proof is fp64 on a small model, which is the appropriate place to establish that the estimator computes the derivative it claims to.

(2) Last-layer identity. Because the residual stream gives $h_{l+1} = h_l + f(h_l)$, the fitted Jacobian must approach identity toward the target, and on our fits $\|J_l - I\|_F/\sqrt{d}$ decreases toward it in every model: Gemma-2-9B $1.0283 \to 0.2706$ (Spearman $\rho = -0.995$), Gemma-2-2B $1.1466 \to 0.4732$ ($\rho = -0.864$), Qwen2.5-7B $1.78 \to 0.594$. The decrease toward the target layer (Figure 4) is overall but not step-by-step: 0.917 of adjacent steps decrease on Gemma-2-9B and 0.625 on Gemma-2-2B, which is what a fitted (rather than analytically derived) Jacobian should be expected to give. We note the metric scale explicitly, because it is easy to misread: $\|I\|_F = \sqrt{d}$, so the zero matrix scores exactly 1.0 under this normalization, and the invariant under test is therefore the *decrease* toward the target, not the absolute level. Controls in the same artifact confirm that the gate detects mis-indexing, mis-scaling, and degeneracy. It is invariant to transposition (the Frobenius norm is transpose-invariant and I is symmetric), so transposition is a failure mode this gate cannot see; it is covered instead by the finite-difference gate, which is sensitive to it.

(3) Positive control. Signature (a) must rise toward the output if the instrument can see present signal, and it does on every model (Gemma-2-9B $0.0102 \to 0.4086$; Qwen2.5-7B, model-argmax convention, $0.004123 \to 0.529514$), which makes the localization interpretable rather than an artifact of a blind instrument. The gate scores signature (a) under the *model-argmax* convention (agreement with the model's own top-1 prediction), whereas the signature profiles reported in Section 5 and Table 2 score against the *actual* next token. The two conventions give the same shape at different levels, and we never combine them within a single table or figure.

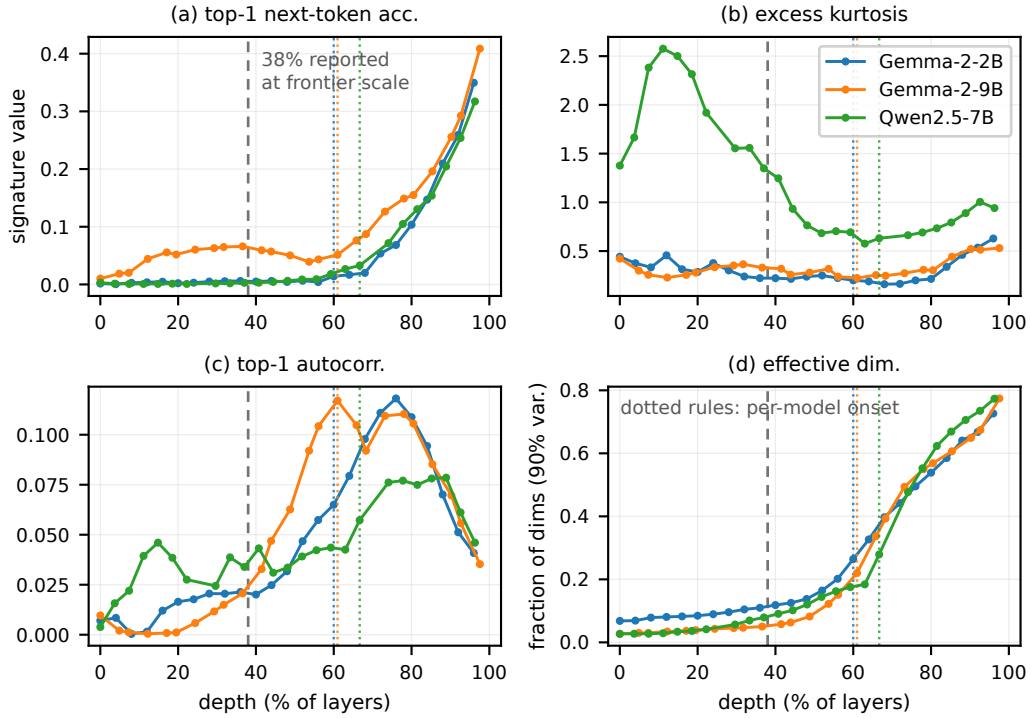

Figure 5: The original account's four band signatures across depth, computed with the fitted averaged-Jacobian lens (actual-next-token convention for (a)). The grey dashed rule marks the 38% reported at frontier scale, drawn as a plain labelled reference line rather than as a shaded band under test; coloured dotted rules mark each model's own onset under R-onset. Signatures (a) top-1 next-token accuracy, (c) top-1 autocorrelation and (d) effective dimensionality organize around those onsets, though (c) is bimodal on Qwen2.5-7B (Appendix E gives the values behind that): (a) rises past them to a motor maximum, (c) is an inverted-U peaking between them and the output, (d) fans out from them. Signature (b), excess kurtosis, does not under the tested fits, and Section 5.1 shows why it cannot serve as a locator there. Plotted directly from the same artifacts audited in the text.

C THE CROSS-FAMILY ONSET LADDER

Signature (d) needs no forward passes, so wherever a lens has been published we can read an onset at the same rule, threshold and depth-resolution convention as our own fits. Table 3 is every rung on the ladder. Its families are Gemma-2 (Gemma Team, 2024), Gemma-3 (Gemma Team, 2025), Llama-3 (Grattafiori et al., 2024), OLMo-3 (Team Olmo, 2025), Qwen3 (Qwen Team, 2025), Qwen3.5 (Qwen Team, 2026a) and Qwen3.6 (Qwen Team, 2026b), with gpt-oss-20B (OpenAI, 2025) among the rungs the resolution screen excludes. No rung is re-cut for signatures and none enters a match count; each carries the onset alone.

Qwen2.5-7B. It carries the signature and capacity work but is not on the ladder, which was merged from the two published-lens ladders and the two large fits of our scale extension; its onset of 66.7% falls inside the reported range, so its inclusion would not change it.

The screen. Four rungs are shaded in Table 3 and dropped, taking the range from 0.0–87.0% to 28.6–78.7%. Gemma-3-270M is read at 17 depths (5.9 points) and is separately degenerate: its (d) curve falls monotonically from layer 0, so min–max normalization puts its maximum at depth zero and the rule returns 0.0% at every threshold in the grid. That is the absence of an onset, not an early one, and it is the reason the unscreened range begins at 0.0%. Qwen3.5-0.8B, Qwen3.5-2B-Base and gpt-oss-20B are each read at 23 depths (4.3 points). Dropping gpt-oss-20B costs the pool its highest onset, 87.0%; we drop it under the criterion rather than keep a number the criterion excludes.

Table 2: Signatures re-cut against each model's own band (onset from R-onset, endpoint at 92%). All three rows use our own fits; the Gemma-2-2B lens is our n=150 fit. Numbers are region means over layers, except "max" which is the argmax over all measured depths (actual-next-token convention for (a) throughout). Occupancy and variance-excess exist for two of these three fits; the Gemma-2-9B capacity row is deferred to Table 5 because its artifact uses a comparison lens.

Model (fit)	Signature	sensory	band	motor	max	at depth
Gemma-2-2B (n=150) band 60.0–92.0%	(a) top-1 acc.	0.00408	0.0991	0.3496	0.3496	96.0%
	(b) exc. kurtosis	0.2994	0.2733	0.6286	0.6286	96.0%
	(c) top-1 autocorr.	0.0205	0.0884	0.0409	0.1181	76.0%
	(d) eff. dim.	0.1074	0.4842	0.7266	0.7266	96.0%
Gemma-2-9B (n=464) band 61.0–90.2%	(a) top-1 acc.	0.0469	0.1372	0.3506	0.4086	97.6%
	(b) exc. kurtosis	0.3023	0.3212	0.5224	0.5306	97.6%
	(c) top-1 autocorr.	0.0271	0.0993	0.0455	0.1170	61.0%
	(d) eff. dim.	0.0572	0.4772	0.7242	0.7743	97.6%
Qwen2.5-7B (n=464) band 66.7–88.9%	(a) top-1 acc.	0.0054	0.1163	0.2856	0.3173	96.3%
	(b) exc. kurtosis	1.4587	0.7335	0.9729	2.5765	11.1%
	(c) top-1 autocorr.	0.0333	0.0737	0.0536	0.0785	88.9%
	(d) eff. dim.	0.0827	0.5512	0.7542	0.7734	96.3%

Table 3: All 26 rungs under R-onset. *step* is the nominal depth points per measured step, 100 over the number of measured depths; actual steps vary about it because samples land on integer layers; shaded rows fail the 4.2-point resolution screen of Section 4 and are excluded from every range in the paper. *sweep* is the width of the rung's onset excursion across the threshold grid $\{0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.50\}$, reported for every rung and not only for the excluded ones (Appendix J). *src*: "our fit" is one of the five lenses of Table 1; "ckpt" is a published mid-fit checkpoint. *flag*: M = weights read through an ungated full-precision mirror rather than the gated original release; it = instruction-tuned. Fit sizes for published lenses are read from each release's metadata and are claimed rather than verified; the Qwen3-32B row shows the two can disagree. *cont. onset* is the continuous locator of Appendix K with its 95% interval; R-onset stays the primary reading.

Model	family	params	depths	step	lens n	src	flag	onset	sweep	cont. onset [95% CI]
Gemma-2-2B	Gemma2	2.60	25	4.00	–	our fit		60.0	24.0	59.0 [57.8, 60.2]
Gemma-2-9B	Gemma2	9.20	25	4.00	–	our fit		61.0	19.5	59.3 [57.7, 60.8]
Gemma-2-27B	Gemma2	27.20	25	4.00	780	published		53.3	15.6	48.8 [43.8, 53.8]
Gemma-2-27B	Gemma2	27.20	25	4.00	464	our fit	M	53.3	13.3	48.6 [44.2, 53.0]
Gemma-3-270M	Gemma3	0.27	17	5.88	302	published	M	0.0	0.0	13.7 [9.4, 18.0]
Gemma-3-1B	Gemma3	1.00	25	4.00	467	published	M	44.0	68.0	56.4 [51.8, 60.9]
Gemma-3-4B	Gemma3	4.30	25	4.00	616	published	M	63.6	48.5	79.2 [56.3, 97.0]
Gemma-3-12B	Gemma3	12.20	25	4.00	775	published	M	57.4	36.2	74.5 [42.1, 97.9]
Gemma-3-27B	Gemma3	27.40	25	4.00	670	published	M	78.7	37.7	91.3 [42.4, 98.4]
Llama-3.1-8B	Llama3	8.03	25	4.00	539	published	M	45.2	19.4	42.4 [39.8, 44.9]
Llama-3.3-70B-It	Llama3	70.55	25	4.00	125	published	Mit	40.5	12.7	36.2 [32.8, 39.5]
OLMo-3-7B	OLMo3	7.30	25	4.00	616	published		35.5	29.0	23.3 [16.5, 30.1]
OLMo-3-32B	OLMo3	32.23	25	4.00	470	published		28.6	7.9	25.5 [23.4, 27.7]
Qwen3-1.7B	Qwen3	2.03	25	4.00	466	published		59.3	37.0	55.7 [54.2, 57.3]
Qwen3-4B	Qwen3	4.02	25	4.00	479	published		60.0	20.0	56.9 [55.3, 58.4]
Qwen3-8B	Qwen3	8.19	25	4.00	461	published		57.1	17.1	55.7 [54.2, 57.1]
Qwen3-14B	Qwen3	14.77	25	4.00	615	published		61.5	20.5	55.6 [54.4, 56.9]
Qwen3-32B	Qwen3	32.76	25	4.00	80	ckpt		77.8	12.7	73.4 [72.4, 74.3]
Qwen3-32B	Qwen3	32.76	25	4.00	464	our fit		73.0	12.7	72.4 [71.5, 73.3]
Qwen3.5-0.8B	Qwen3.5	0.87	23	4.35	233	published		56.5	30.4	52.3 [49.8, 54.8]
Qwen3.5-2B-Base	Qwen3.5	2.27	23	4.35	335	published		52.2	30.4	50.7 [48.7, 52.7]
Qwen3.5-4B	Qwen3.5	4.66	25	4.00	417	published		51.6	25.8	48.3 [45.6, 51.1]
Qwen3.5-9B-Base	Qwen3.5	9.65	25	4.00	458	published		48.4	32.3	45.1 [42.8, 47.3]
Qwen3.5-27B	Qwen3.5	27.78	25	4.00	672	published		69.8	28.6	67.6 [64.5, 70.7]
Qwen3.6-27B	Qwen3.6	27.78	25	4.00	1000	published		69.8	23.8	67.2 [65.1, 69.3]
gpt-oss-20B	gpt-oss	20.90	23	4.35	457	published	it	87.0	4.3	86.9 [86.5, 87.3]

Fit size is a second axis the resolution screen does not test. The published Qwen3-32B artifact is not a finished lens but a mid-fit checkpoint: it holds an unnormalized `jacobian_sum` with `n_done` = 80 and none of the `J`, `d_model` or `n_prompts` fields every other rung ships, while its own configuration record reports 615 prompts fitted, contradicting the artifact. Normalizing by n is harmless for signature (d), whose value is a ratio of eigenvalues and therefore invariant to a

global scale on J, so the checkpoint can be read; what cannot be repaired is that it is fit on 80 prompts. Llama-3.3-70B-Instruct is fit on 125. Every other rung is fit on 417 to 1000. Excluding both leaves 20 rungs and seven families and does not move the range, which stays at 28.6–78.7%. For Qwen3-32B the checkpoint is superseded by our own $n{=}464$ fit, which returns 73.0% against the checkpoint's 77.8%.

Tuning is not uniform. Llama-3.3-70B-Instruct and gpt-oss-20B are instruction-tuned and are marked as such. The Qwen3.5 ladder is the harder case: two of its five rungs are explicitly base checkpoints (Qwen3.5-2B-Base, Qwen3.5-9B-Base) and three carry no such marking. Under the naming convention the Qwen3 rungs follow, an unmarked name denotes the post-trained model, so this ladder most likely mixes base and instruction-tuned rungs; we have not been able to confirm that against the release cards, and we flag it rather than assume either way. No claim in this paper rests on a Qwen3.5 rung individually, but three of them sit inside the 22, so the mix is a property of the range and is stated here for that reason.

Architecture is not uniform inside Gemma-3. The Gemma-3 onsets run 44.0% (1B), 63.6% (4B), 57.4% (12B) and 78.7% (27B), a swing of 34.7 points, the largest within-family movement anywhere in the ladder. It spans an architecture change. The 1B rung is a text-only stack with a $262{,}144$-entry vocabulary whose unembedding is read at `model.embed_tokens`; the 4B rung and above place the text stack under a multimodal wrapper, carry $262{,}208$ entries, and are read at `language_model.model.embed_tokens`. The step from 1B to 4B therefore changes scale, vocabulary and unembedding path at once. We report the swing and attribute it to none of the three.

Weights provenance. Two repository families are license-gated to the account used for this work, so their weights were read through ungated full-precision mirrors: the five Gemma-3 rungs and both Llama-3 rungs, marked M in Table 3, and additionally our own Gemma-2-27B fit, which was fit against mirrored weights and is marked in Table 1. Eight of the 26 rungs are affected. Only the unembedding and the final-norm gain are read from these weights, and the mirrors were not compared tensor-by-tensor against the original releases, so this is a provenance caveat we carry rather than one we have discharged. Two of the eight, Gemma-3-27B at 78.7% and Llama-3.3-70B-Instruct at 40.5%, sit near opposite ends of the range.

Sweep widths. The width of each rung's threshold sweep is an observation we record without a trend claim and without an explanation. The Gemma-3 sweeps are wide: 68.0 points at 1B and 48.5 at 4B, against 13.3 to 24.0 for the Gemma-2 rungs. It is not a grid artifact: the Gemma-3 rungs share the 25-depth grid with the Gemma-2 rungs. The sweep column of Table 3 gives each rung's width. The two families differ in tokenizer and in architecture as well as in generation, and reading one family's signature curve against the other's would compare spectra computed over different-sized vocabularies through different unembeddings, which is the cross-architecture step we decline elsewhere in this appendix. The practical consequence is only that a Gemma-3 onset is a weaker point estimate than a Gemma-2 one, and should be read with the sweep column beside it.

D THE BOUNDARY CUT SEALED IN ADVANCE

A locating rule we announced and then abandoned should be reported rather than inferred, and it bounds how much of the band's *extent* we can claim. Section 3.3 reports why the sealed pre-motor rule (the layer before the largest single-step increase in c_{90}) was not used. For completeness, this is what it returns. On Gemma-2-2B it puts the boundary at 64.0%, giving a two-layer band; on Gemma-2-9B our wikitext fit puts it at 61.0% and our reasoning fit at 92.7%; on Qwen2.5-7B it returns 66.7%, coinciding with the onset and leaving a single-layer band. Under that cut signature (a) scores on 6 of 6 profiles and (d) on 6 of 6, while (c) drops to 2 of 6 and (b) to 2 of 6, the drop in (c) being an artifact of bands too narrow to contain a maximum rather than a change in the underlying profile, which is the same profile plotted in Figure 5. The near-tie behind that instability is the reason: on Gemma-2-9B the two candidate steps in c_{90}, at layer 27 (65.9% depth) and layer 40 (97.6%), are separated by 0.0025 in a curve of total range 0.747, our wikitext fit giving 0.1169 against 0.1002, our reasoning fit 0.1102 against 0.1166, and a comparison lens 0.1197 against 0.1222. Both cuts are written to the same artifact by the analysis script, so a reader can recompute either.

E LOCATORS THAT DO NOT USE SIGNATURE (D)

Section 4 varied the model–lens artifacts under one rule. This section varies the rule on fixed models, and the two must not be run together: artifact-disagreement and rule-disagreement are separate failures of the instrument, and merging them attributes to the models a spread that belongs to the locator. R-onset reads the effective-dimensionality transition off (d), and Section 5 then scores (a) and (c) against the resulting operational edge, so part of their agreement is built in. To separate the two we derive an onset from (c) and from (a) alone, by rules fixed before their outcomes were inspected and using the same normalization and 0.25 threshold as R-onset, so nothing favours (d): (c)-x25 and (a)-x25, the first depth at which the normalized signature exceeds 0.25 of its range, and (c)-peak, the depth of (c)'s inverted-U maximum (Table 4).

Two of the three agree with (d) in direction and disagree in magnitude. Signature (a) gives 80.0, 73.2 and 77.8%: a 6.8-point spread, as tight as (d)'s own, but 14.4 points *later* on average. Signature (c)'s peak gives 76.0, 61.0 and 88.9%, a wider 27.9-point spread. Neither touches (d). That is the part of the circularity objection we can answer: on these three models the onset is not an artifact of anchoring on (d). It says nothing about the artifacts from the six other families of Section 4, which no rule but R-onset has been applied to.

The third locator fails. Signature (c)'s first crossing gives 48.0% on Gemma-2-2B and 41.5% on Gemma-2-9B but 11.1% on Qwen2.5-7B: the only onset any of the four rules places in the sensory region on any of these three models, and the largest single disagreement in Table 4. The cause is visible in Figure 5(c): Qwen's autocorrelation is bimodal, with an early local maximum of 0.0460 at 14.8% depth before the later rise to 0.0785 at 88.9%, and a first-crossing rule fires on the early bump. Such a rule presumes a monotone rise, which (c) does not have; we fixed it in advance and report it as it came out.

A consensus rule, and why it does not rescue the comparison. The audited account locates its band where its several analyses identify a similar range of layers. That is a qualitative procedure and its per-layer values are unpublished, so it cannot be reproduced, only reimplemented; everything in this paragraph is our arithmetic rendering of it and licenses no statement about what their procedure would return. We fixed the rendering before running it: min–max normalize each of the four signatures, take each one's *half-rise*, the first depth reaching 0.5 of its range, and call the median of the four the consensus onset. The half-rise is deliberately not R-onset's 0.25 crossing, so the two rules differ in kind and not only in level.

It does not return a systematically earlier onset. On Gemma-2-2B it gives 66.0% against R-onset's 60.0%, on Gemma-2-9B 61.0% against 61.0%, and on Qwen2.5-7B 44.4% against 66.7%: later, identical, then earlier. We had sealed in advance that a consensus rule returning an onset at least 15 points earlier on all three fits would make the gap against the frontier-scale figure locator-driven. It does not.

The reason is visible in the same numbers and is the more useful finding. Within a single model the four half-rises do not agree at all: they span 88.0 points on Gemma-2-2B, 90.2 on Gemma-2-9B and 85.2 on Qwen2.5-7B, which is nearly the whole depth of each model. Signature (b), early-peaked on these fits, puts its half-rise at 0.0, 0.0 and 3.7%; (a) puts its at 88.0, 90.2 and 88.9%. A median over four landmarks that far apart is not a consensus, it is a choice of which four to take, and moving any one of them moves it. On these three fits there is no narrow range for several analyses to agree on, and that is a statement about the instruments rather than about the models.

F SIGNATURE (B): THE TWO DIAGNOSTICS AND THE FULL TRIMMING CURVE

Excess kurtosis matches on 3 of 6 profiles, and the three that match are all Gemma-2-9B. That pattern is not evidence of a scale effect; it is what a measurement looks like when it is dominated by a few extreme values. Two diagnostics establish this, and both are properties of the instrument rather than of any model.

Instability across fits. Across two of our own Gemma-2-9B fits (n=464 each, differing only in fit corpus), Pearson r over the 25 measured depths is 0.9897 for (a), 0.9845 for (c) and 0.9996 for (d),

Table 4: Band onset under four locating rules, all fixed before their outcomes were inspected and all using the same normalization and 0.25 threshold. Only (d)-x25 is used to define the band elsewhere in the paper; the other three are independent of it. The bottom row is the disagreement among the four rules *within* each model, which is the quantity this table exists to show. Bold marks the one rule–model combination that dissents from the other three, and Section 4.1 explains why the rule that produces it fires early.

Model (fit)	(d)-x25	(a)-x25	(c)-peak	(c)-x25	rule spread
Gemma-2-2B (n=150)	60.0%	80.0%	76.0%	48.0%	32.0
Gemma-2-9B (n=464)	61.0%	73.2%	61.0%	41.5%	31.7
Qwen2.5-7B (n=464)	66.7%	77.8%	88.9%	11.1%	77.8
spread across models	6.7	6.8	27.9	36.9	

but -0.1731 for (b). Two fits of one model at one size, differing only in the corpus the Jacobian was averaged over, disagree about the sign of its depth profile.

Outlier domination. Section 5.1 reports the sign flip at $M = 10$: removing the M most extreme vocabulary entries, activations held fixed so that only the dictionary changes, moves the band-minus-sensory contrast on our own lens, against the band as reported at frontier scale, from -0.0031 at $M = 0$ to $+0.0066$ at $M = 10$. The mechanism is structural: (a) and (c) are rank statistics and (d) is spectral, all bounded in their sensitivity to single vocabulary entries, whereas (b) is a fourth central moment over a $256{,}000$-way distribution and is not. The rest of the curve points the same way. As M goes from 0 to 1000, the mean of (b) over depth falls from 0.3246 to -0.0285: the entire kurtosis level lives in under half a percent of the $256{,}000$-entry vocabulary. Over the same range, agreement between the two lenses rises monotonically, from $r = 0.9285$ at $M = 0$ to 0.9704 at $M = 1000$, which is what an outlier-driven disagreement looks like rather than a structural one. Under the (d)-anchored band the band-minus-sensory contrast does not flip (it runs $+0.0408$, $+0.0462$, $+0.0559$, $+0.0583$ at $M = 0, 10, 100, 1000$), so the sign instability is specific to the window the contrast is taken over, which is itself a reason not to use (b) to choose a window. Wherever (b) is reported we suggest a trimmed column alongside the raw one.

Trimming repairs (b) across five independent fits, at one trim level. The diagnostics above compare two lenses. To ask whether trimming is a *repair* rather than a description of one pair, we fitted five further Gemma-2-2B lenses for this purpose, each n=150 at the frozen regime: three disjoint wikitext-103 prompt shards and two disjoint GSM8K shards, so three differ from one another only in which prompts the Jacobian was averaged over and two differ in corpus as well. All five are read through identical activations, so the profiles differ solely in J. We fixed the bar in advance: every one of the ten pairwise Pearson correlations over the 25 depths at or above 0.90, *and* no sign change in the band-minus-sensory contrast across the five. Both conditions, or the claim is not made.

Untrimmed, both fail. The minimum pairwise correlation is 0.8616, and the contrast runs $+0.0060$, -0.0155, $+0.0075$, $+0.0450$ and $+0.0588$: one of the five fits puts it on the opposite side of zero. That is the instability Section 5.1 reports, reproduced across independent fits rather than across two lenses of one fit.

At $M = 1000$ both conditions hold. The minimum pairwise correlation rises to 0.9190 and every contrast is positive, at $+0.0256$, $+0.0191$, $+0.0298$, $+0.0421$ and $+0.0484$. Removing 1000 of $256{,}000$ vocabulary entries, under half a percent, is what decides the sign. On these five fits a trimmed excess kurtosis is reproducible where the raw statistic is not, which is the repair the trimming curve above suggests and this tests.

Two qualifications belong with that verdict. The rise is not monotone in M: the minimum correlation is 0.8616, 0.8561, 0.8576 and 0.9190 at $M = 0, 10, 100, 1000$, so agreement rises end to end but dips first, and only the largest trim on the sealed grid qualifies. And 0.9190 clears a 0.90 bar by 0.019 on five fits, which is a pass rather than a comfortable one. We report the repair at the trim level where it holds and do not claim it holds at a smaller one.

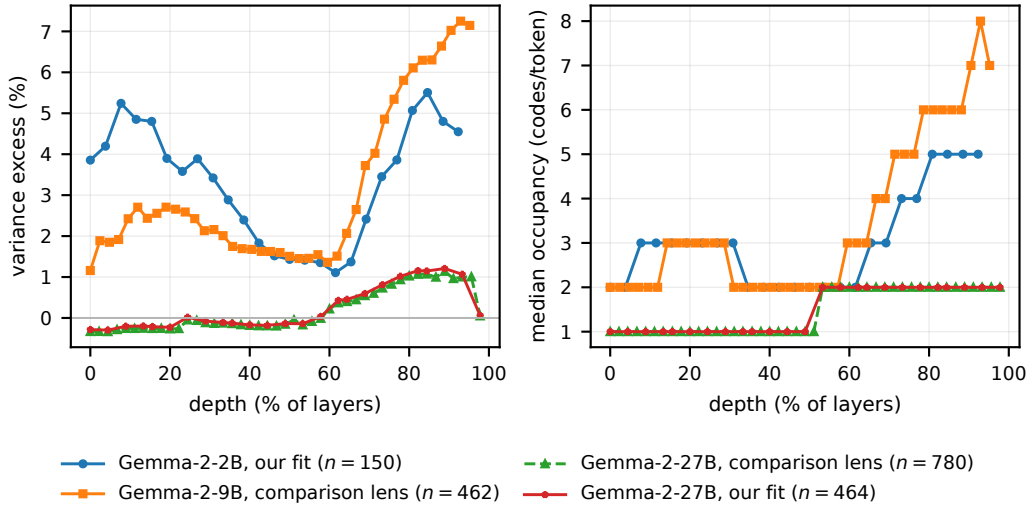

Figure 6: Left: variance excess against depth on three Gemma-2 scales. The two Gemma-2-27B curves are the same model read through two different lenses, our own n=464 fit and an independently published n=780 lens, and they lie almost on top of each other, so the 27B reading is not a property of one lens. Right: median occupancy, flat at 1–2 at 27B under both lenses and rising late at 2B and 9B without leaving single digits under the uncentered unembedding. Note that the three scales are not sampled on the same grid: Gemma-2-9B is measured at 41 depths and the published-lens 27B curve at 45, against 25 for Gemma-2-2B and our own 27B fit, so the 9B maximum is drawn from a finer grid and the three maxima are not strictly comparable.

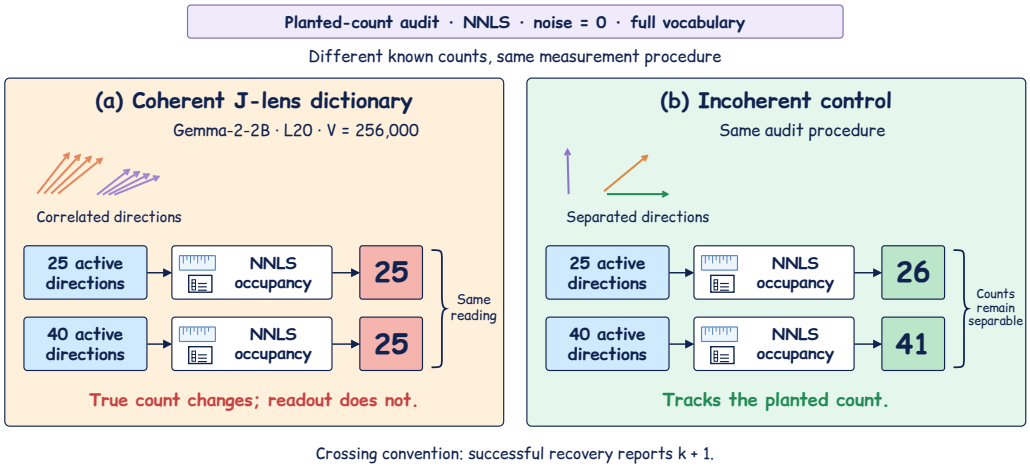

Figure 7: Why a repeated occupancy value cannot establish a capacity limit. On the coherent Gemma-2-2B L20 dictionary at full vocabulary, non-negative least squares returns the same reading at 25 and at 40 planted directions; on an incoherent control the same procedure separates them. Successful recovery reports $k + 1$ under the crossing convention of Appendix G.

G Occupancy under both unembedding conventions, and the sparse-recovery regime

G.1 Where these dictionaries sit relative to the classical conditions

Occupancy is a sparse-recovery problem: it decomposes an activation against an overcomplete dictionary and reports how many atoms the decomposition uses. That problem has a classical theory,

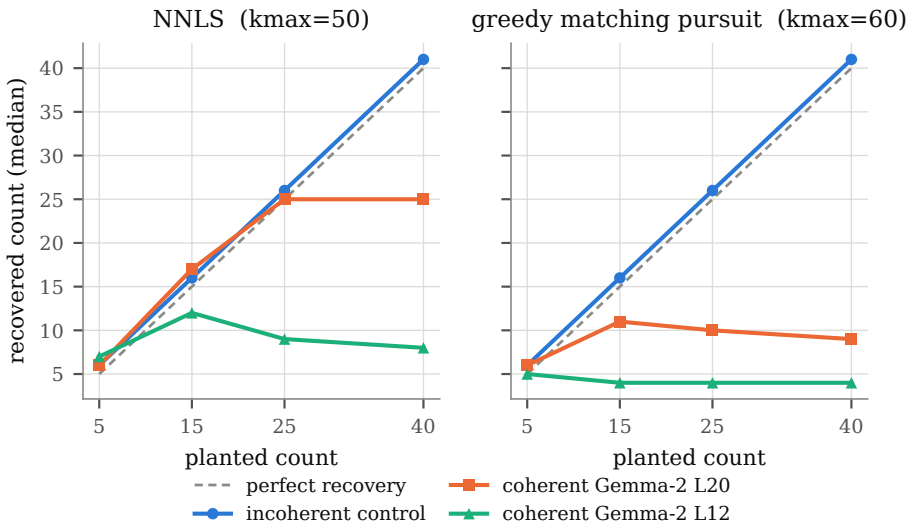

Figure 8: Occupancy-operator saturation on synthetic ground truth (noise 0.0). On an incoherent control dictionary both operators return exactly one more than the planted count: the crossing is read at the first K that fails to beat the random control, so successful recovery reads $c{+}1$. On the coherent Gemma-2 J-lens dictionaries both fall below it at the higher planted counts, and under greedy matching pursuit Gemma-2 L12 is flat from a planted count of 15 onward. NNLS dominates greedy matching pursuit at every planted count *empirically*, and still saturates. This is why the tested occupancy estimators cannot test the capacity claim under the U-raw unembedding; Section 5.2 reports the U-centered result.

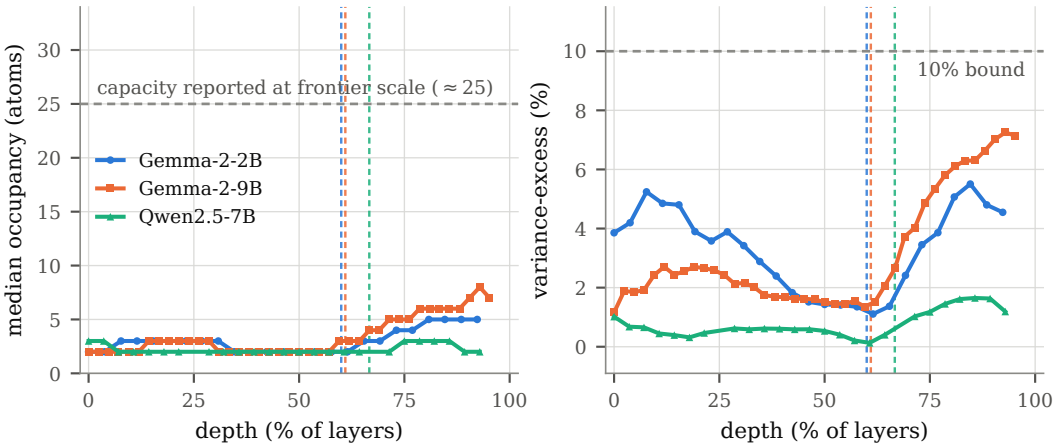

Figure 9: Left: median occupancy against depth, with the capacity reported at frontier scale as a reference line. A band would appear as an elevated plateau between each model's onset (coloured rules) and the output. Under the uncentered unembedding the estimator never leaves single digits on Gemma-2 or on Qwen2.5-7B: no model approaches the reference line at any depth, so there is no elevated plateau to read a band off. Centered, it does leave them (Section 5.2). Right: variance-excess, which keeps responding, against the reported 10% bound. The Gemma-2-9B curve in both panels comes from a comparison lens (Table 5).

and the theory should be stated before any of our own measurements, because it already says these dictionaries lie outside the regime in which such a decomposition is guaranteed to recover anything.

Two results fix the regime, both in terms of the mutual coherence μ, the largest absolute inner product between two distinct unit-norm atoms. Donoho & Elad (2003) show that a K-sparse representation is the unique sparsest one whenever $K < \mathrm{spark}(D)/2$, where $\mathrm{spark}(D)$ is the smallest

number of linearly dependent atoms, and give the computable relaxation $\mathrm{spark}(D) \geq 1 + 1/\mu$. Tropp (2004) turns the same quantity into a condition on the algorithms actually used, showing that greedy orthogonal matching pursuit and basis pursuit both recover the true K-sparse representation whenever

$$K \;<\; \tfrac{1}{2}\,(1 + 1/\mu)\,, \tag{1}$$

equivalently whenever $\mu < 1/(2K-1)$. Reading a capacity of 25 off such a decomposition therefore asks for $\mu < 0.0204$, and telling 25 from 40 asks for $\mu < 0.0127$.

The J-lens dictionaries are nowhere near either figure. On Gemma-2-2B the mean coherence is $\bar{\mu} = 0.343$ at L12 and 0.108 at L20, and the maximum coherence, which is the quantity Equation 1 is stated in, is 1.000 at both depths: the vocabulary contains atom pairs that are collinear to measurement precision. Equation 1 then permits $K < 1.00$ at either depth, and the spark bound certifies uniqueness for no sparsity level at all. Substituting $\bar{\mu}$, which is not what the theorems are stated in but is the more forgiving reading, permits $K < 1.96$ at L12 and $K < 5.12$ at L20. Every route through the classical conditions lands at least a factor of about five below the count the capacity claim turns on.

What this does and does not establish. These conditions are *sufficient*, not necessary: failing them withdraws the guarantee, but does not prove that recovery fails and does not predict *which* dictionaries fail. Our own incoherent control settles the second point: its maximum coherence of 0.127 licenses only $K < 4.45$, so it violates the condition at 25 exactly as the real dictionaries do, and it still recovers exactly one more than each planted count through the same code. No dictionary on which we compute it satisfies Equation 1 at $K = 25$; the incoherent control and the synthetic dictionaries up to α=1 in Appendix H recover it anyway. The classical theory therefore fixes a fact about the experiment rather than about its outcome. The theory withdraws the guarantee, and Section 5.2 reports what happens when the check is run: recovery fails on five of the six dictionaries under the U-raw unembedding and succeeds on all six under the U-centered one. We compute the condition only on the U-raw dictionaries, so it cannot say which construction recovers, which is why the check is empirical.

What is inherited and what is new. Three claims adjacent to this one are already in the literature and we do not restate them as findings. The first two are the target account's own: that the J-lens frame is overcomplete, so an activation admits many decompositions rather than one (their §2.3), and that occupancy counts active J-lens directions rather than concepts or ideas (their §4.2). Their §2.3 further notes that the sparsity level at which the decomposition is taken is itself a choice rather than a measurement. The third is Wang & Reid (2026), who reimplement the same occupancy estimator on looped and depth-recurrent architectures and report that readout claims can be estimator-specific, comparing their reimplementation against a second readout and reporting where the two agree and where they disagree. Those are claims that a number moves: with the decomposition, with the concept–direction gap, with the estimator. Ours is adjacent: on synthetic data where the active-direction count is known by construction, the operator fails to separate two different true counts under one unembedding convention and separates them under the other (Appendix G.5). Three things here are ours and are not implied by the above: the planted-count demonstration itself, the location of the failure on a coherence axis (Appendix H), and the negative result that the real dictionaries do not fail where coherence-matched synthetics do. In the other direction, Ivanitskiy et al. (2026) give the positive theory this failure sits against, including an exact recovery threshold in terms of coherence for real equiangular tight frames. It bounds a different decoder, a thresholded correlation rule rather than the pursuit and non-negative least squares we run, and its sharp results are stated for tight and equiangular dictionaries, which these are not; we cite it as the statement of what identification would require here, not as a prediction of our outcome.

G.2 The synthetic grid

The grid plants a known number of active directions, $\{5, 15, 25, 40\}$, at noise 0.0, and asks each occupancy operator to recover that count by decomposing the resulting activation against a dictionary. Three dictionaries are used: the coherent Gemma-2 J-lens dictionaries at L12 and L20, and an incoherent control. This grid decomposes against the *full* $256{,}000$-entry vocabulary, which is the regime the capacity claim concerns; the coherence sweep of Appendix H instead holds $V = 20{,}000$

so that real and synthetic dictionaries are matched, and the two differ for NNLS on L20 for exactly that reason.

Figure 8 shows the resulting recovery curves. Non-negative least squares (NNLS; Lawson & Hanson, 1974) recovers $6/17/25/25$ on L20 and $7/12/9/8$ on L12; greedy matching pursuit (Mallat & Zhang, 1993) recovers $6/11/10/9$ on L20 and $5/4/4/4$ on L12; the incoherent control recovers $6/16/26/41$, exactly one more than the planted count at each of the four counts, confirming that the failure is a property of dictionary coherence rather than of the harness. That offset is not approximation error. Occupancy as we implement it from the published description reports the first K at which the real marginal gain falls below the matched random control, and that is one greater than the number of atoms which beat it, so a dictionary on which recovery succeeds reads $c{+}1$ at a planted count of c. The offset is constant wherever recovery succeeds, absent wherever it fails, and leaves every rank statistic we report unchanged. We describe the original account's estimator only as its text describes it: its implementation was not released, and we make no claim about what its reported capacity would be under this convention.

Two features of the grid are worth isolating. First, the coherent-dictionary curves are not merely biased but *flat* over part of the range: under greedy matching pursuit Gemma-2 L12 returns the same recovered count, 4, from a planted count of 15 onward, and NNLS on L12 falls from 12 to 8 as the planted count rises, so a measured occupancy in that regime carries no information about the true count. Second, the ordering between operators is stable (NNLS is $\geq$ greedy at every planted count, *empirically*, on every dictionary), so the saturation is not an artifact of choosing a weak operator; at full vocabulary the better of the two saturates as well, returning 25 at both 25 and 40 planted directions on L20.

One property of the crossing estimator should be recorded. A curve that never crosses is reported at $k_{\max}$ rather than as missing, so a censored cell cannot be distinguished from a genuine crossing at $k_{\max}$. Across the $\sim 15{,}600$ per-token occupancy cells behind every occupancy number in this paper, exactly one sits at its run's $k_{\max}$: a single token of 48 in one synthetic cell of the coherence sweep ($\alpha{=}1.25$, NNLS, planted 40), and that cell may itself be a genuine crossing. Treating it as unresolved instead leaves the median of its cell unchanged and moves the corresponding 25-vs-40 separability from 0.786 to 0.784, on the same side of the 0.75 bar. No cell on either Gemma-2 dictionary, and none in the operator grid of Figure 8, sits at $k_{\max}$. We report the value as computed.

G.3 Occupancy read on the real models

Read on our own fits for Gemma-2-2B and Qwen2.5-7B, and on a comparison lens for Gemma-2-9B (Table 5), all under the uncentered unembedding, occupancy is floored on every dictionary and none follows the predicted band profile. Median occupancy spans 2 to 5 over the full depth of Gemma-2-2B, 2 to 8 on Gemma-2-9B and 2 to 3 on Qwen2.5-7B, against a reference capacity of 25, and 2–5, 3–6 and 2–3 inside the respective bands. An estimator that never leaves single digits cannot exhibit the elevated in-band plateau the band profile predicts, whichever region one cuts. Centered, the same sweep reads 6 to 19 over the depth of Gemma-2-2B and 6 to 10 inside its band, and 16 to 30.5 on Qwen2.5-7B, 20 to 25 inside its band.

The unembedding convention, and why it must be fixed. Every occupancy and variance-excess number in this paper other than those marked centered is built from an uncentered unembedding, and the choice is not inert; Section 5.2 states what it decides. Logits change only by a constant under that shift, so a readout cannot tell the two conventions apart, but a decomposition can. `transformer_lens` (Nanda & Bloom, 2022) applies `center_unembed` by default and silently declines for models with a logit soft-cap, which Gemma-2 has and Qwen2.5 does not, so one unchanged line of code yields an uncentered dictionary for one family and a centered one for the other. Measured here, the two Qwen constructions differ by 29% in Frobenius norm, while the two Gemma constructions are bit-identical. The centered Qwen dictionary is close to orthogonal, and under it occupancy reads near the frontier-scale capacity where the uncentered dictionary reports it floored like the others. The centered-construction artifact is retained alongside the reported one.

Table 5: Capacity measures inside each model's band, with lens provenance stated. The Gemma-2-9B row is computed from a comparison lens rather than one of our own fits: the full-depth occupancy sweep for that model was not recomputed on our n=464 fit, and is marked as such. The two rows that back the reference-free claims are the Gemma-2-2B and Qwen2.5-7B rows.

Model	lens	occ. (all depths)	occ. (band)	var-exc. (band)	at depth
Gemma-2-2B	our fit, n=150	2–5	2–5	5.51%	84.6%
Gemma-2-9B	comparison lens	2–8	3–6	6.64%	88.1%
Qwen2.5-7B	our fit, n=464	2–3	2–3	1.65%	85.7%

G.4 The variance-excess denominator, and what rescaling cannot fix

Variance-excess is the measure Section 5.4 reports in occupancy's place, so the convention it is computed under is stated here explicitly. It is a ratio, and the origin its denominator is taken about has to be the origin its numerator is taken about. The numerator is the residual energy of a decomposition; if that decomposition reconstructs the raw activation, the denominator must be the raw energy of the same vectors, and if the activations are centered before decomposing, it must be the centered variance. Both conventions are self-consistent. Mixing them is not: a residual of raw activations divided by a centered total is not a fraction of anything, is not bounded by one, and is inflated by the ratio of raw to centered energy. The two conventions are used for different quantities, and they concern different objects. *Activation-centering* is what this paragraph is about; *U-centering*, the subtraction of the mean vocabulary row from the unembedding, is a separate choice treated in Appendix G.3. Every real-model occupancy and variance-excess number (Section 5.4, Appendix G.3, Figures 6 and 9, Table 5) is activation-centered: the activations are centered before they are decomposed, and the denominator is the centered variance of the same matrix. The planted-count gate of Appendix G.5 is activation-raw. Its planted activations are not centered, the operators reconstruct raw h_ℓ (Table 8, row 6), and its variance-excess divides by the raw energy of the same vectors. Within each route, numerator and denominator agree.

The reason to be explicit is that the inflation is not a nuisance constant that a later rescaling could absorb, and the way it fails to be one is asymmetric. The factor depends only on the activation set, not on the number of directions K at which the measure is read. Anything defined as a fraction of the measure's own maximum over K is therefore invariant to it: a knee located as the smallest K reaching a fixed fraction of the peak is unchanged, because a positive rescaling moves the peak and the threshold together. Anything that compares the measure across *different* activation sets is not invariant, because those sets have different geometries and therefore different factors. Near-parallel activations lose most of their energy to centering and are inflated heavily; near-orthogonal activations lose almost none and are inflated barely at all. A rank correlation of variance-excess against a varying active-direction count, a difference between two counts in percentage points, and an area under a curve separating two counts all move under the mismatch, and they move by different amounts at different counts.

The practical consequence: a reading taken under a mismatched denominator cannot be repaired after the fact by dividing through, and has to be recomputed. No single constant recovers it, because no single constant was applied. This also says which diagnostics are robust to the choice and which are not: shape-of-curve statistics survive, cross-condition comparisons do not, and it is why the coherence of a dictionary matters here for a second reason, beyond the recovery failure of Appendix G.2: the same code is nearly exact on an incoherent dictionary and badly wrong on a coherent one, so an incoherent control cannot certify it.

G.5 A planted-count gate, sealed in advance, on the operator the capacity claim uses

Appendix G.2 tests two operators, and neither is the one the original account names. This subsection reports an extension that closes that gap and two others. It adds non-negative gradient pursuit, the operator the capacity claim is actually built on. It densifies the planted-count grid from the four levels of Figure 8 to seventeen, $\{1, 2, 3, 4, 5, 6, 8, 10, 12, 15, 18, 21, 25, 30, 35, 40, 50\}$, at 10 seeds and 256 tokens per cell, so that a resolution limit can be estimated rather than asserted. And it

puts variance-excess, the measure Section 5.4 offers in occupancy's place, through the same gate that removed what it replaces. Every threshold quoted below was fixed in writing before the run it judges, together with the sentence each outcome would license; the specification, three amendments, a correction and an addendum are versioned with the code, and the two places where the eventual result did not fit the sentence written for it are marked below.

An independent implementation. No implementation of the original estimator was released. Ours is written from the published description and from the algorithm that description cites, base gradient pursuit (Blumensath & Davies, 2008), made non-negative by projecting the coefficients after each update: unit-normalised atoms, selection on strictly positive correlations only, one new atom per step, the published exact line-search step, and the raw residual-stream activation as the reconstruction target. We implemented it without consulting any reimplementation, so that agreement or disagreement with the original account's numbers is evidence rather than construction. It is a first measurement and not a confirmation of one: we have run no other gate on this operator that left a stored result, and we claim no second implementation behind it.

The harness is gated on an incoherent control. The sealed plan blocks every real-dictionary result unless the harness first tracks planted counts on an incoherent control dictionary. It does. All three operators return exactly one more than the planted count at all seventeen levels, which is the $c{+}1$ reading that successful recovery produces under this crossing convention, with $\rho(\hat{K}, c) = 1.0000$, $\mathrm{AUC}(25, 40) = 1.000$ and no ceiling.

The operator the original account uses fails in the same way. Under our independent implementation, the recovered count saturates on both Gemma-2 dictionaries and does not separate 25 planted directions from 40 at the sealed bar. Gradient pursuit returns a median of 8 at every planted count from 18 to 50 at L12, and at L20 it rises to 24 at a planted count of 25 and then falls back to 17 by 50; the separability is $\mathrm{AUC}(25, 40) = 0.500$ at L12 and 0.000 at L20, against a bar of 0.75. An AUC of 0.000 is an inverted ordering, not chance: the operator reliably reports fewer directions at 40 planted than at 25. The bar is one-sided because a count estimator that falls as the count rises cannot be read as a count, so such a cell fails it. The failure is therefore not an artifact of testing two operators nobody uses: all three operators we tried, including the one the capacity claim is built on, are unable to distinguish the counts the claim turns on (Table 6). The same holds at the sealed secondary noise level of 0.1: every ceiling in Table 6 is unchanged, one separability moves (NNLS at L20, from 0.500 to 0.400, still far below the bar), and exactly two of the 102 occupancy medians move, from 12 to 11.5 and from 11 to 11.25.

Two incidental checks ride on the same run. First, the four-point figures Appendix G.2 reports were computed with the two operators run at different token counts and different caps; re-run here under identical conditions they come back unchanged, at $6/17/25/25$ and $7/12/9/8$ for NNLS, $6/11/10/9$ and $5/4/4/4$ for greedy matching pursuit, and $6/16/26/41$ on the control, so that mismatch moves no value reported here. Second, the ordering sentence in the same subsection survives the matched re-run: NNLS is $\geq$ greedy at every one of the seventeen planted counts on every dictionary.

Variance-excess under the same gate. Section 5.4 proposes variance-excess where occupancy failed, and the same planted-count grid can ask whether the replacement survives the test that removed what it replaces. It can only ask that of a statistic that is in range, and the numerator and denominator must share a convention. The asymmetry Appendix G.4 sets out is why: a mismatched pair is nearly exact on the incoherent control, where centered energy is 99.6% of raw energy, and wrong by more than an order of magnitude at L12, where centered energy is 7.42% of raw energy at a planted count of 25 (25.44% at L20). An incoherent control therefore cannot certify the statistic.

Two gates were sealed ahead of the re-run and both were passed before it started. The first recomputed the incoherent control under the corrected raw denominator and required the premise Amendment 01 rests on to survive, which it did: $\rho(\mathrm{VE}, c) = -1.0000$ for all three operators, every cell in range. The second is a positive control on coherent geometry, where the two denominators differ by more than an order of magnitude and the true value of the statistic is known independently of the operators, run across four arms that vary how far the planted activations depart from orthogonality, with the first reproducing the incoherent Phase 1 geometry (median centered-energy

Table 6: The densified planted-count gate at noise 0.0, 10 seeds, full $V = 256{,}000$. $\hat{K}$ is the median recovered count. $R(c)$ is the resolution at c: the smallest grid step d for which $\mathrm{AUC}(c, c{+}d) \geq 0.75$, and NaN if no grid point above c reaches the bar. C is the ceiling: the smallest planted count above which nothing separates, NaN if there is none. $R(50)$ is NaN in every cell, the control included, because 50 is the top of the grid and no pair $(50, c')$ exists to evaluate; it is a boundary of the design and not a measurement of saturation. Gradient pursuit is the operator the original account names.

		$\hat{K}$ at planted count					$R(c)$			
dictionary	operator	5	15	25	40	50	c=5	c=25	AUC(25, 40)	C
incoherent	greedy MP	6	16	26	41	51	1	5	1.000	NaN
control	NNLS	6	16	26	41	51	1	5	1.000	NaN
	gradient pursuit	6	16	26	41	51	1	5	1.000	NaN
Gemma-2	greedy MP	5	4	4	4	4	NaN	NaN	0.500	4
L12	NNLS	7	12	9	8	8	1	NaN	0.000	10
	gradient pursuit	8	10	8	8	8	1	NaN	0.500	10
Gemma-2	greedy MP	6	11	10	9	8	1	NaN	0.000	12
L20	NNLS	6	17	25	25	21	1	5	0.500	30
	gradient pursuit	6	17	24	21	17	1	NaN	0.000	25

fraction 0.99229 down to 0.98993 across the grid) and the last an order of magnitude below it. Its six sealed predictions all held: the corrected statistic is in $[-1, 1]$ in all 1,020 cells and the defective one exceeds 1 in all 255 cells of each coherent arm; the algebraic identity relating the two holds to 2.2×10^{-16}; and across the four arms the corrected statistic varies by at most 0.0156 at any planted count, where the defective one varies by 31.9 at a planted count of 1 alone.

Under the corrected statistic the amended criterion is met, on both Gemma-2 dictionaries, under all three operators, which is the outcome that keeps Section 5.4's claim. Variance-excess moves monotonically with the planted count ($\rho = -1.0000$ on five of the six dictionary-by-operator cells and -0.9975 on the sixth; the criterion is two-sided, and we print the sign because every cell we measure is negative), it moves by 29.04 to 55.26 percentage points across the grid rather than drifting, and it separates 25 planted directions from 40 in every cell at the maximum two-sided value of 1.000. The non-saturation Section 5.4 claims is therefore a property we have checked against known ground truth, not one observed only on real activations where no ground truth exists. The response is comfortably above the 2.0 percentage-point floor sealed against an arbitrarily small but perfectly ordered drift, and above the 4.3 percentage points that separate the 27B in-band maximum from the 2B one, which is the scale at which this paper treats a difference as real elsewhere.

Two qualifications belong with that verdict and both were required in advance. *The sign is negative.* Variance-excess measured at $K = c$ *falls* as more directions are planted, in all six cells, from 0.9909 at a planted count of 1 to between 0.4382 and 0.7005 at 50. A monotone response is the property the criterion tests and the property Section 5.4 needs, and the criterion was made two-sided before any Gemma-2 data existed for exactly this reason. But a negative response is not evidence that variance-excess is a count estimator, and it must not be read as one: variance-excess does not track the planted count, it responds monotonically to it. *And no ceiling was found.* The sealed wording for this outcome ends by attaching a saturation point to the measure, on the expectation that one would be found on the grid. None was: the ceiling is NaN in all six cells. It is not established as unlimited either: no saturation point appears on this grid, so what we can report is that variance-excess is still separating at 50 planted directions, not that it separates without limit. That departs from the sealed sentence, which assumed an outcome that did not occur, and the departure is recorded in the amendment file.

The exploratory count estimator sealed alongside the primary, the smallest K at which variance-excess reaches 95% of its maximum, is ours and is used by neither this paper nor the original account, and nothing here depends on it. It behaves like occupancy rather than like variance-excess: on the control its median rises at every one of the seventeen levels, reaching 44 at a planted count of 50, while under gradient pursuit on the real dictionaries it saturates at 3 from a planted count of 18 at L12 and turns over after 14 at L20. The one sealed secondary that was not run is a variant of gradient pursuit carrying the coefficient gate Wang & Reid (2026) operate the same estimator under,

Table 7: The sealed planted-count gate on six J-lens dictionaries: the two Gemma-2-2B dictionaries of Table 6 and four added here, at in-band depths chosen by a rule fixed before the runs. $\mathrm{AUC}(25, 40)$ is the separation between 25 and 40 planted directions, against a sealed bar of 0.75, under greedy matching pursuit (GMP), NNLS and non-negative gradient pursuit (GP). Left, the unembedding as loaded for Gemma-2, uncentered: one cell of six separates. Right, W_U with its vocabulary mean subtracted, the construction `transformer_lens` applies by default to Qwen2.5 but not to soft-capped Gemma-2: all six separate. All twelve cells use the same harness, bars, grid and ten seeds.

		uncentered W_U				centered W_U			
dictionary	depth	GMP	NNLS	GP	sep.?	GMP	NNLS	GP	sep.?
Gemma-2-2B L12	48.0%	0.500	0.000	0.500	no	1.000	1.000	1.000	yes
Gemma-2-2B L20	80.0%	0.000	0.500	0.000	no	1.000	1.000	1.000	yes
Gemma-2-9B L25	61.0%	0.500	0.000	0.000	no	1.000	1.000	1.000	yes
Gemma-2-9B L37	90.2%	0.000	1.000	1.000	yes	1.000	1.000	1.000	yes
Qwen2.5-7B L18	66.7%	0.400	0.000	0.100	no	1.000	1.000	1.000	yes
Qwen2.5-7B L24	88.9%	0.050	0.000	0.000	no	1.000	1.000	1.000	yes

which admits an atom only if its coefficient exceeds $0.05\|h\|$; it is sealed and unreported, and no result here rests on it.

Where resolution runs out, and where the grid does. The two quantities in Table 6 say how fine a distinction each operator can still make. On the control, resolution is a single grid step up to a planted count of 5, widens to 5 at 25 and to 10 at 40, and no ceiling is reached: the operator is still separating at the top of the grid. On the real dictionaries it runs out early. $R(c)$ is NaN from a planted count of 10 onward for both NNLS and gradient pursuit at L12, from 4 onward for greedy matching pursuit at L12, and from 12 onward for greedy matching pursuit at L20; the ceilings are 4, 10 and 10 at L12 and 12, 30 and 25 at L20. The best cell in the table is NNLS at L20, which still resolves at a planted count of 25 and resolves nothing from 30 onward. One reading has to be excluded explicitly: $R(50)$ is NaN everywhere, including on the control, and that is the grid ending rather than the operator stopping. There is no grid point above 50 to compare 50 against.

The gate on a second family and on in-band depths, and where it stops holding. The grid above is two depths of one model. We extended it, under the same sealed harness and the same bars, to two in-band depths each of Gemma-2-9B and Qwen2.5-7B, chosen by a rule fixed in advance: the first sampled source layer at or after that model's own R-onset, and the last at or below the adopted 92% endpoint. Ten seeds, full vocabulary, noise 0.0. No cell was censored and none returned a NaN occupancy: 0 of 43,520 per-token cells sit at $k_{\max}$ in every one of the four new dictionaries.

Under the uncentered unembedding, non-identification holds on five of the six dictionaries we have now tested, and fails to hold on the sixth. Gemma-2-2B at L12 and L20, Gemma-2-9B at its onset layer L25, and Qwen2.5-7B at L18 and L24 all return $\mathrm{AUC}(25, 40)$ at or below 0.5 under all three operators, the operator the audited account names included. Gemma-2-9B at L37, at 90.2% depth, does not: NNLS and non-negative gradient pursuit both separate the two planted counts at $\mathrm{AUC} = 1.000$, NNLS tracking the planted count to a median of 27 at 25 and 36 at 40 with no ceiling anywhere on the grid.

L37 was then the only near-output Gemma-2 dictionary we had gated, and it read as a lone counterexample. Gating every sampled layer at 85–95% depth (Appendix G.6) makes it the first point of a pattern: every Gemma-2-2B dictionary from 88.0% and every Gemma-2-9B dictionary from 85.4% separates, and no Qwen2.5-7B dictionary does at any depth we test. Occupancy's non-identification is a property of the dictionary, its depth and its family, not of J-lens dictionaries as a class. Under the centered unembedding it disappears: the gate separates the two counts on all six dictionaries, each tracking the planted count as the incoherent control does (Section 5.2). Centering removes a direction every row shares, and planted activations keep 0.98–1.00 of their energy about their mean, against 0.05–0.31 uncentered. Table 7 gives every cell under both conventions.

An ordering, sealed before the last two cells existed. The one dictionary that separates is also the one whose fitted Jacobian is closest to identity. We sealed that prediction while the two Qwen dictionaries were still running, with the falsifier written down: Qwen2.5-7B L18, at $\|J_\ell - I\|_F/\sqrt{d} = 1.2385$ the largest of the six, had to *fail* or the ordering was refuted. It fails. Sorted by distance the six run 0.6013 (L37, separates), 0.6898, 0.8947, 0.9452, 1.1299, 1.2385, with no failing dictionary closer to identity than the separating one (on the twelve dictionaries of Appendix G.6 the ordering interleaves), and Spearman $\rho = -0.3339$ on $n = 6$, quoted descriptively and without a p-value.

We state what that is worth, which is not much on its own. One dictionary of six separates, so "no interleaving" is a condition a single correctly-placed point satisfies, and it would occur about one time in six by chance. Distance from identity is also collinear with depth across these six, by construction, since the fitted Jacobian approaches identity toward the target layer; nothing here separates the two. Six dictionaries from two families is descriptive. The ordering is reported for what it is: where the lens is close to the unembedding, the dictionary it induces is less coherent and recovery is easier, and a quantity the validation gate already computes says which regime a dictionary is in.

The reproducibility cost of an unreleased estimator. The published description does not determine the estimator. Table 8 lists every point at which we had to choose, with the choice and its reason, because that list is the cost of reimplementing a measure whose code was not released. One of the eleven is marked: the original says nothing about how the random control set is matched, and that choice bears directly on the occupancy value rather than on the algorithm that produces it. Row 9 is left unmarked for the reason its own cell gives, that the units question is inert for the crossing. One further row is not the original account's underdetermination at all but our own invention, and is labelled accordingly.

What this arm does not establish. The arm is synthetic throughout. Planted activations are constructed as non-negative sparse combinations of dictionary atoms plus noise, and real residual-stream states are not known to be generated that way; a failure to recover a planted count bounds the operator on data of the planted kind, and establishes nothing about what it does on activations whose true generative structure is unknown. Planted sparse-linear structure is moreover the assumption most favourable to the estimator, and if real activations are less sparse-linear in this dictionary the operator's position is worse rather than better, but that is an argument and we do not make it as a measurement. The real dictionaries are the six of Table 7, from three models in two families, so nothing here characterises other families or the frontier-scale model the original claim was measured on. And a negative result about an instrument is not a negative result about the claim: showing that occupancy cannot separate 25 from 40 does not show that the workspace holds some other number of directions, nor that it holds none. One limitation of the coherent control should be recorded with it, since it was found after the gate passed and bears on how the gate may be cited. Of its six predictions, five are independent; the sixth, that the two statistics agree on the incoherent arm, is very nearly implied by the manipulation check that precedes it, because the algebraic identity between the statistics makes the quantity it measures a function of the same centered-energy fraction that check is defined on. It came out as predicted.

G.6 The gate at 85–95% depth

The one dictionary that separated under the uncentered unembedding, Gemma-2-9B at L37, sits near the output, where the J-lens approaches the logit lens. To ask whether the estimator becomes identifiable above some depth, we gated every sampled source layer of our three gated fits at 85–95% depth that had not been gated: six further dictionaries, under the same harness, bars, grid and ten seeds, noise 0.0, and the uncentered unembedding, the construction under which non-identification holds and a depth effect could therefore show. The layers, the decision rule and the wording for each outcome were sealed before the runs.

Table 9 gives all twelve uncentered cells. Within each family the separating dictionaries are all deeper than the failing ones, and across families no one depth divides them, so the sealed outcome is the family-dependent one: identifiability varies with depth and family, without a threshold common to the families we test. Gemma-2-2B separates from somewhere in $(80.0, 88.0]\%$, Gemma-2-9B from somewhere in $(61.0, 85.4]\%$, and Qwen2.5-7B at no depth we test up to 92.6%. The Gemma-

Table 8: Degrees of freedom in reimplementing the occupancy estimator from its published description. † marks the choice the original account leaves entirely unspecified and that most moves the occupancy value itself rather than the algorithm. ‡ marks a definition that is ours, invented to make an exploratory test possible, and that neither this paper nor the original account otherwise uses. All eleven were fixed in writing before any run.

	left underdetermined	resolved as
1	Non-negativity is asserted, but the cited algorithm is unconstrained	Project coefficients to $\max(\cdot, 0)$ after each update: the smallest change that enforces the stated constraint and leaves the published step rule intact
2	Sign handling in selection under non-negativity	Select the largest strictly positive correlation; a negatively correlated atom cannot enter a non-negative combination
3	Atom normalisation is not stated	Unit-normalise dictionary columns, else selection ranks norms rather than correlations
4	Whether an atom may be re-selected	No; the support grows by exactly one per step, which is what makes K the step index the occupancy definition needs
5	Stopping, beyond the cap k	Stop at $k_{\max}$, when no positive correlation remains, or when the exact-step denominator vanishes; the last two are degeneracy guards the original does not mention
6	Decomposition target: raw h_ℓ or a centered basis	Raw h_ℓ, which is what the original says it reconstructs
7	Step size	The published exact line search
8	Which variant of the cited algorithm	Base gradient pursuit, not its conjugate variants; the original names the algorithm without qualification
9	"Marginal improvement in reconstruction" in what units	Change in reconstruction error per added atom, compared real against random on the same activation. For the crossing itself this is inert: explained variance and squared error differ by a per-activation constant shared by both curves, so the crossing K is identical either way. It binds on aggregation, and we aggregate over per-activation crossings, never over raw deltas
10†	How the random control set is matched	Same V, d, column norms, cap and code path, drawn i.i.d. Gaussian and unit-normalised, one draw per seed. The original says "random directions of the same size" and nothing more. This is the choice that most moves the occupancy value
11‡	A count estimator from variance-excess, which neither paper defines	Smallest K at which variance-excess reaches 95% of its maximum over $K \in [1, k_{\max}]$. Ours, exploratory, load-bearing for nothing; the 95% is arbitrary within a broad range

2-9B interval is wide because no Gemma-2-9B layer between those depths was gated, not because the boundary is uncertain there by measurement; it bounds where the boundary lies and does not locate it.

Three details qualify the pattern. Greedy matching pursuit separates on none of the twelve uncentered dictionaries; every Gemma-2 separation comes from NNLS and non-negative gradient pursuit. The weakest separation is Gemma-2-2B at L22, which separates under NNLS alone, with medians of 27 at 25 planted directions and 29 at 40, and should not be read as equivalent to the others. And the ordering by distance from identity that held on six dictionaries does not hold on twelve: Qwen2.5-7B at L25 fails at 0.6402, closer to identity than three separating Gemma-2 dictionaries, and Spearman ρ between distance and the best operator's $\mathrm{AUC}(25, 40)$ is -0.5137 on $n = 12$ against -0.3339 on the first six, quoted descriptively. Under the centered unembedding none of this arises: all six dictionaries gated centered separate (Table 7).

Table 9: The uncentered planted-count gate on twelve J-lens dictionaries: the six of Table 7 and six added at 85–95% depth (†). $\|J_\ell - I\|$ is $\|J_\ell - I\|_F/\sqrt{d}$. AUC$(25, 40)$ against a sealed bar of 0.75 under greedy matching pursuit (GMP), NNLS and non-negative gradient pursuit (GP). *Separates under NNLS alone, with medians of 27 and 29 at 25 and 40 planted directions.

dictionary	depth	$\|J_\ell - I\|$	GMP	NNLS	GP	separates?
Gemma-2-2B L12	48.0%	1.1299	0.500	0.000	0.500	no
Gemma-2-2B L20	80.0%	0.9452	0.000	0.500	0.000	no
Gemma-2-2B L22†	88.0%	0.7933	0.000	1.000	0.650	yes*
Gemma-2-2B L23†	92.0%	0.6884	0.000	1.000	1.000	yes
Gemma-2-9B L25	61.0%	0.8947	0.500	0.000	0.000	no
Gemma-2-9B L35†	85.4%	0.6805	0.000	1.000	1.000	yes
Gemma-2-9B L37	90.2%	0.6013	0.000	1.000	1.000	yes
Gemma-2-9B L38†	92.7%	0.5500	0.000	1.000	1.000	yes
Qwen2.5-7B L18	66.7%	1.2385	0.400	0.000	0.100	no
Qwen2.5-7B L23†	85.2%	0.7725	0.300	0.000	0.000	no
Qwen2.5-7B L24	88.9%	0.6898	0.050	0.000	0.000	no
Qwen2.5-7B L25†	92.6%	0.6402	0.500	0.000	0.000	no

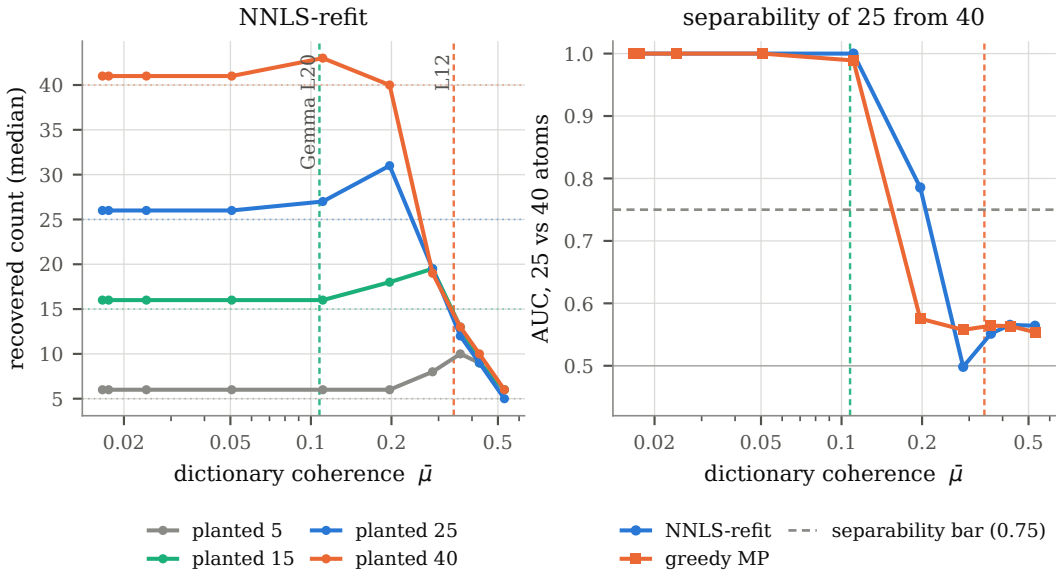

Figure 10: Occupancy non-identification against dictionary coherence. Left: recovered count collapses onto a single curve as coherence rises, so the planted counts stop being separable. Right: the 25-vs-40 rank statistic from which the threshold is read, with the separability bar at 0.75 (dashed) and chance at 0.5 (solid). In the left panel, which is NNLS-refit, dotted lines mark the planted counts 5, 15, 25 and 40; under the crossing convention exact recovery sits one unit above its line. Vertical rules mark the two real Gemma-2 J-lens dictionaries. Those rules are the point of the figure: the real dictionaries do not sit where their mean coherence says they should (Section H.3).

H THE COHERENCE SWEEP, AND WHAT IT LOCALIZES

Appendix G places the J-lens dictionaries outside the classical recovery conditions and demonstrates that recovery fails there. Where on a coherence axis the failure sets in is mapped below, with one negative result that matters for how the failure should be described: the real dictionaries do not fail where their coherence statistics say a synthetic dictionary of the same coherence would.

Table 10: Coherence sweep: median recovered count at each planted count, and the 25-vs-40 separability AUC, for both operators. The synthetic family degrades monotonically with $\bar{\mu}$; the two real Gemma-2 dictionaries do not sit where their $\bar{\mu}$ predicts. Gemma-2 L20 is the sharpest case: it fails under greedy matching pursuit at a mean coherence where the synthetic family still separates perfectly. Vocabulary: this sweep holds $V = 20{,}000$ atoms per dictionary so that synthetic and real dictionaries are matched. That is why NNLS separates on L20 here (AUC 0.982) while the full-vocabulary grid of Appendix G shows it saturating; the dependence is measured in Section H.2. Coherence statistics ($\bar{\mu}$, $\mu_{\max}$) are computed over the full dictionary, as in Table 11.

			greedy MP					NNLS-refit				
dictionary	$\bar{\mu}$	$\mu_{\max}$	5	15	25	40	AUC	5	15	25	40	AUC
α=0	0.017	0.122	6	16	26	41	1.000	6	16	26	41	1.000
α=0.25	0.018	0.124	6	16	26	41	1.000	6	16	26	41	1.000
α=0.5	0.024	0.196	6	16	26	41	1.000	6	16	26	41	1.000
α=0.75	0.050	0.452	6	16	26	41	1.000	6	16	26	41	1.000
α=1	0.110	0.733	7	18	29	43	0.989	6	16	27	43	1.000
α=1.25	0.195	0.891	9	22	26.5	28	0.575	6	18	31	40	0.786
α=1.5	0.283	0.956	12.5	16	15	16	0.558	8	19.5	19.5	19	0.498
α=1.75	0.359	0.980	10.5	11	10.5	11	0.564	10	13	12	13	0.551
α=2	0.423	0.990	8	9	8	9	0.564	9	9	9	10	0.566
α=2.5	0.523	0.997	6	6	5	6	0.554	6	6	5	6	0.564
Gemma-2 L12	0.343	1.000	5	5	4	4	0.505	7	16	12	10	0.202
Gemma-2 L20	0.108	1.000	6	11	11	10	0.246	6	17	27	33	0.982

H.1 The sweep

We build a family of synthetic power-law dictionaries indexed by an exponent α, spanning mean coherence $\bar{\mu}$ from 0.017 to 0.529, and run the two occupancy operators of Appendix G.2, greedy matching pursuit and NNLS, against planted counts $\{5, 15, 25, 40\}$ at noise 0.0, over 3 seeds, with $V = 20000$ atoms in $d = 2304$ dimensions, 48 tokens and a budget of 60 steps. Separability is scored as the AUC for discriminating 25 planted atoms from 40; we call a dictionary *failed* when that AUC falls below 0.75. Table 10 gives the grid and Figure 10 plots it.

Read across the synthetic rows alone, the sweep looks like a clean story: both operators recover the planted count essentially exactly up to $\bar{\mu} \approx 0.11$, then degrade, and the interpolated failure threshold sits at $\bar{\mu}^* = 0.160$ for greedy matching pursuit and 0.208 for NNLS (interpolated on the sweep's own 4000-atom subsample estimate of $\bar{\mu}$, not on the full-dictionary values tabulated above). The real dictionaries break that story, which is what the rest of this appendix is about.

H.2 Two controls

Vocabulary size is not the confound. The real dictionaries are subsampled to $V = 20000$ of 256000 vocabulary entries, so a natural worry is that failure is an artifact of how densely we sample the vocabulary. Sweeping V through 20000, 100000 and 256000 raises near-neighbour density sharply (on L20 the fraction of atoms with a neighbour above cosine 0.7 goes from 0.079 to 0.271 to 0.438) while greedy separability stays flat at AUC 0.213, 0.188 and 0.199, and the recovered counts barely move ($11.5/10$, $11/9$, $10/9$ for 25 and 40 planted). On L12 greedy is likewise flat (0.443, 0.503, 0.453). Greedy failure is therefore not caused by vocabulary size. One exception, stated: NNLS on L20 *does* depend on V, falling from AUC 0.993 at $V = 20000$ to 0.686 and then 0.433. So the L20 row of Table 10 that shows NNLS separating is a small-vocabulary result and does not survive to the full vocabulary; the V-invariance claim is a claim about greedy matching pursuit only.

The harness is not the cause. Appendix G already reports the incoherent control recovering $6/16/26/41$, exactly one more than each planted count, through the same code path, so the failure is a property of the dictionary rather than of the implementation. We do not repeat that grid here.

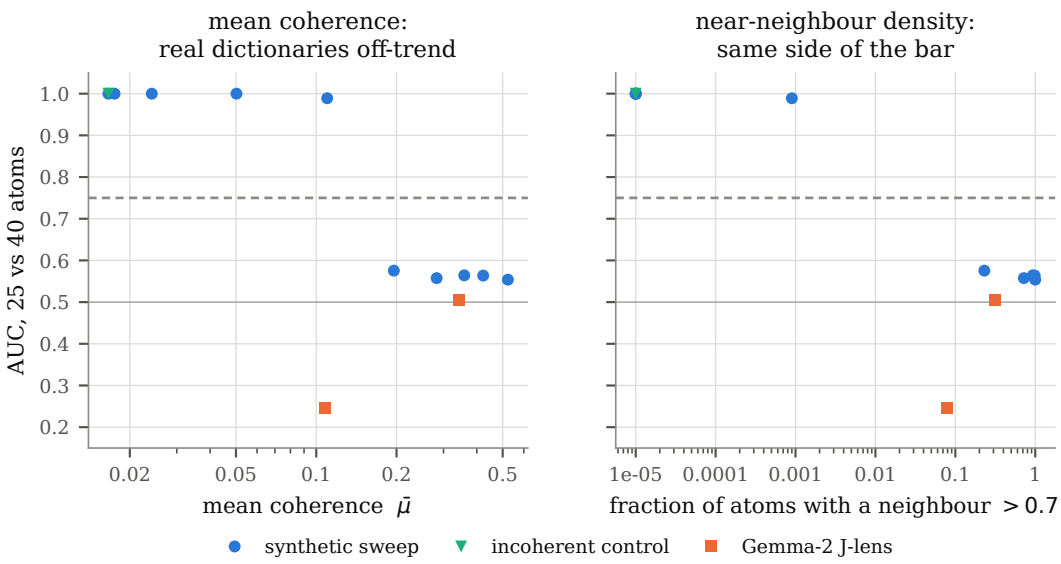

Figure 11: Why mean coherence is the wrong statistic. Left: separability against $\bar{\mu}$; the synthetic family traces a clean threshold and the real J-lens dictionaries sit far off it, failing at a mean coherence where the synthetic family still separates. Right: the same points against near-neighbour density, which puts real and synthetic failures on the same side of the bar. The dashed line is that bar, AUC 0.75, and the solid line is chance, 0.5. The incoherent control and the synthetic dictionaries up to α=0.75 have no neighbour above 0.7; their exact zeros are drawn at the axis floor, 10^{-5}. The panel does *not* collapse them onto one curve: Gemma-2 L20 still fails harder than any synthetic dictionary at its density, and the panel is not drawn to suggest otherwise.

H.3 Coherence predicts the regime, not the individual dictionary

Equation 1 is a worst-case condition: it is stated in $\mu_{\max}$ and it is violated by every dictionary here, the ones that recover as well as the ones that do not. So it cannot be read as a graded predictor, and the sweep makes that concrete. A matched pair is the cleanest demonstration. Gemma-2 L20 and the synthetic α=1 dictionary have the same ambient dimension ($d = 2304$) and effectively the same mean coherence (0.108 against 0.110), yet L20 fails at AUC 0.246 while α=1 passes at 0.989. Nor does the discrepancy live in the coherence tail: Table 11 shows the 99th, 99.9th and 99.99th percentiles all *lower* for the failing dictionary than for the passing one (ratios 0.70, 0.78, 0.91), so every quantile orders the pair the wrong way. Across the 13 dictionaries with matched ambient dimension, of which 7 fail, only three statistics separate failure from success with zero errors: $\mu_{\max}$, and near-neighbour density at cosine 0.5 and 0.7. Mean coherence makes one error, and so does every percentile statistic we tried.

Three graded accounts that do not hold. Three graded accounts of the failure are ruled out. *Mean coherence* is refuted by the matched pair just given. *Near-duplicate atoms* are not necessary for failure: the synthetic α=1.25 dictionary fails while containing no atom pair above cosine 0.9 at all, and adding such pairs does not graduate the failure. A *geometric crossing argument* (that the real and random marginal-gain curves are both geometric and cross at a step where the true support size does not appear) requires the residual-to-atom cosine ϕ_k to plateau, and it does not: over the interval the crossing depends on, ϕ_k varies by a factor of 4.69 to 5.91 on the real Gemma-2 dictionaries, while only the random curve is flat (ratio at most 1.34). Its predicted crossings miss the measured occupancies by a median of 146.1 steps and are anti-correlated with them, and on all four real cells it is undefined outright. All three were the wrong question. The right one is prior to them and is answered in Appendix G: whether the recovery problem is well-posed on these dictionaries at all. It is not, and that is why no statistic of the dictionary's pairwise geometry grades the outcome.

One candidate survives in weakened form and we note it as a direction rather than a result. The *effective rank of the planted support* (the idea that the operator returns the dimension the planted

Table 11: The matched pair, and which statistic separates. Left block: values for the failing (Gemma-2 L20, AUC 0.246) and passing (α=1, AUC 0.989) dictionary, matched at $d = 2304$. A ratio below 1 means the statistic orders the pair the wrong way. Right column: errors made when the statistic is thresholded to separate failure from success across all 13 matched-dimension dictionaries. Only $\mu_{\max}$ and near-neighbour density at 0.5 and 0.7 make none.

statistic	fail (L20)	pass (α=1)	ratio	separation errors
mean coherence $\bar{\mu}$	0.108	0.110	0.99	1
$\mu_{\max}$	1.000	0.733	1.36	0
coherence p99	0.260	0.373	0.70	1
coherence p99.9	0.370	0.476	0.78	1
coherence p99.99	0.506	0.553	0.91	1
nbr. density > 0.3	0.889	0.937	0.95	1
nbr. density > 0.5	0.355	0.194	1.83	0
nbr. density > 0.7	0.079	0.001	88.28	0
nbr. density > 0.9	0.003	0	n/a	1
largest cluster @0.9	37	1	37.00	1
edges @0.9	330	0	n/a	1
largest cluster @0.99	4	1	4.00	1
edges @0.99	6	0	n/a	1

atoms actually span, not how many were planted) predicts measured occupancy with a median absolute error of 3.0 steps on the synthetic cells but 19.0 on the four real Gemma-2 cells, so it does not work where it would matter. No claim in this paper rests on it.

H.4 What we are reporting

The saturation is robust. It reproduces under both operators of the sweep and all three of the gate (Appendix G.5), across 3 seeds, at zero noise, with an incoherent control that tracks the planted count through the same code, and, for greedy matching pursuit, independently of how much of the vocabulary is sampled. It is also, as Appendix G sets out, the expected regime: these dictionaries fail the sufficient conditions for sparse recovery by a wide margin. Every coherence statistic in this appendix is computed on the U-raw dictionaries. What remains open is why the real U-raw dictionaries fail harder than synthetic dictionaries matched to them on every pairwise coherence statistic we can compute; we leave it open. The claim the main text rests on is narrower: under the U-raw unembedding, occupancy cannot separate 25 from 40 on five of the six dictionaries we gate (Appendix G.5), and under the U-centered unembedding it separates them on all six.

I What carries the depth profile: a decomposition inside one model

Signature (d) is the effective dimensionality of the centered dictionary $W_U^\top J_\ell$. Two matrices, one of which does not vary with depth, so the driver is separable inside a single model in a way the cross-family ladder cannot be. Four tests separate them, sealed in full before any was run.

The object, and why four tests are the four tests. Writing $A := W_U^\top$ for the vocabulary-by-d dictionary factor and $\bar{w}$ for the mean row of A, the centered Gram the statistic reads is

$$A_c \;=\; J_\ell^\top \big(W_U W_U^\top - |V|\,\bar{w}\bar{w}^\top\big) J_\ell \;=\; J_\ell^\top M J_\ell, \tag{2}$$

with M the centered unembedding Gram, $d \times d$ and independent of ℓ. Signature (d) therefore depends on the unembedding *only* through M, and on depth only through J_ℓ. T1 removes the depth variation, T2 removes M, T3 holds both spectra fixed and rotates M's eigenbasis against J_ℓ, and T4 asks where J_ℓ leaves the regime in which T1's answer holds.

Scope, which is narrow. Three lenses, two families. T1–T3 need the stored per-layer Jacobians, and only Gemma-2-2B (n=150), Gemma-2-9B (n=464) and Qwen2.5-7B (n=464) have them;

the Gemma-2-27B and Qwen3-32B Jacobians were not retained and cannot be refit on the storage available. The exclusion is storage, not lens provenance. What follows is a statement about what carries the depth profile in three models, not about the 22-artifact ladder and not about why its model–lens artifacts differ.

T1, the identity control, which is an algebraic near-certainty. Setting $J_\ell = I$ makes Equation 2 read $A_c = M$ at every layer, a matrix with no ℓ in it, so the profile is exactly constant and R-onset, which min–max-normalizes it, is undefined. Run as a harness check rather than as evidence, it returns exactly that: the spread of $\mathrm{frac90}$ across depth is 0 to machine representation on all three lenses, the participation ratio holds at 0.123084, 0.077970 and 0.162931, and R-onset is NaN. We report it as the control it is. It establishes only that the entire depth profile enters through J_ℓ, which is what makes T2 interpretable.

T2, the Jacobian alone. Effective dimensionality of J_ℓ by itself, no unembedding. The statistic is an *analogue* rather than the same measurement, and we say so: J_ℓ is $d \times d$ where the dictionary is $|V| \times d$, so its d rows are treated as atoms and centered over rows by the same code. Its onset lands on the same grid index as the dictionary onset on all three lenses, zero steps apart: 60.0% on Gemma-2-2B, 61.0% on Gemma-2-9B, 66.7% on Qwen2.5-7B.

The restatement objection, and what answers it. If the effective rank of J_ℓ simply tracks (d), T2 is a restatement rather than a mechanism, and the correlations say it nearly is: the participation ratio of J_ℓ's squared singular values correlates with $\mathrm{frac90}$ at $r = 0.9423$, 0.9472 and 0.8951 across 25 depths, and the T2 statistic itself at $r \geq 0.9991$. Those are descriptive, quoted with their n and without a p-value. What distinguishes a restatement from a mechanism is not a tighter correlation but T3, and only T3: under a restatement reading only spectra enter, so rotating M's eigenbasis while holding both spectra exactly fixed must leave the profile alone. T2 varies J_ℓ across depth and holds nothing fixed, so it cannot separate a spectrum that broadened from an alignment that changed; in a real model both move together with ℓ. T3 is the only manipulation here that holds J_ℓ bit-identical, holds both spectra bit-identical, and varies the alignment alone.

T3, and a disagreement between two dimensionality statistics. Replacing M by $Q\Lambda Q^\top$ for Haar-random orthogonal Q preserves its spectrum exactly and randomizes its orientation against J_ℓ. Two sealed preconditions gate the reading and both clear: the manipulation is a no-op wherever $J_\ell \approx cI$, and $\|J_\ell - I\|_F/\sqrt{d}$ exceeds the 0.5 bar at $24/25$, $24/25$ and $25/25$ sampled depths; and it is a no-op if M is isotropic, where $\mathrm{pr}(M)/d$ is 0.1231, 0.0780 and 0.1629 against an inapplicability bar of 0.9.

Read through $\mathrm{frac90}$, the statistic every onset in this paper is read through, the onset is bit-identical across ten Haar draws on all three lenses: spread 0.000, zero index steps from the measured onset. Read through the participation ratio, it is not. On Qwen2.5-7B the onset moves from 66.7% to 74.1%, two grid steps, identically across all ten seeds; on Gemma-2-2B it scatters across 60.0–68.0% between seeds; on Gemma-2-9B it agrees at 61.0%. Whether the onset is invariant to the unembedding's orientation is therefore itself estimator-dependent: the participation ratio shows the disagreement immediately, the same estimator dependence reported elsewhere.

The sealed plan named the participation ratio as a tiebreaker only for the case where $\mathrm{frac90}$ fails to move, which did not occur. No sealed rule covers a disagreement between them, and we do not supply one now, with both answers visible; that would convert an under-specification into a result. We report both and let the disagreement stand as the finding.

What T3's zero spread does and does not mean. It does not mean the curve is unchanged. The randomization moves $\mathrm{frac90}$ by up to 5.556×10^{-2}, 3.990×10^{-2} and 4.632×10^{-2} against a quantisation of 4.340×10^{-4} and 2.790×10^{-4}, so the profile moves by one to two hundred quanta. What does not move is the grid index at which the min–max-normalized curve crosses 0.25. The onset is stable under the manipulation; the curve it is read from is not.

T4, the departure from identity. $\|J_\ell - I\|_F/\sqrt{d}$ is already computed by the validation gate; on Qwen2.5-7B, which that artifact does not carry, it is recomputed from the lens tensor. The onset does not coincide with where J_ℓ stops being near-identity. On Gemma-2-2B the metric first falls

Table 12: The decomposition, three lenses. T1 is a control and returns the constant profile Equation 2 requires, so its onset is undefined. Index steps are on each lens's own 25-point depth grid, whose spacing differs between lenses. T3 is reported under both dimensionality statistics because they disagree; no sealed rule adjudicates between them.

	Gemma-2-2B n=150	Gemma-2-9B n=464	Qwen2.5-7B n=464
d_{model}	2304	3584	3584
R-onset, signature (d)	60.0%	61.0%	66.7%
T1 frac90 spread across depth	0	0	0
T1 R-onset	NaN	NaN	NaN
T2 onset, J_ℓ alone	60.0%	61.0%	66.7%
T2 index steps from (d)	0	0	0
r(eff. rank J_ℓ, (d)), n=25	0.9423	0.9472	0.8951
T3 onset, frac90, 10 seeds	60.0%	61.0%	66.7%
T3 seed spread, frac90	0.000	0.000	0.000
T3 onset, participation ratio	60.0–68.0%	61.0%	74.1%
T3 max $\lvert\Delta\text{frac90}\rvert$	5.556×10^{-2}	3.990×10^{-2}	4.632×10^{-2}
frac90 quantisation $1/d$	4.340×10^{-4}	2.790×10^{-4}	2.790×10^{-4}
Precondition: depths with $\lVert J_\ell - I\rVert_F/\sqrt{d} > 0.5$	24/25	24/25	25/25
Precondition: $\text{pr}(M)/d$ (bar 0.9)	0.1231	0.0780	0.1629
T4 crossing at 0.5	96.0%	97.6%	undefined
T4 index steps from (d)	9	9	NaN

Table 13: T3 in full: R-onset under each Haar draw of Q, all ten sealed seeds, read through both dimensionality statistics. The two agree on Gemma-2-9B and disagree on the other two lenses. Under frac90 the onset is bit-identical across draws everywhere; under the participation ratio it sits two grid steps away on Qwen2.5-7B at every seed, and varies between seeds on Gemma-2-2B. No sealed rule adjudicates between the two statistics, and we do not supply one here.

lens	statistic	seed									
		0	1	2	3	4	5	6	7	8	9
Gemma-2-2B	frac90	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0
n=150	pr	64.0	68.0	68.0	60.0	68.0	60.0	68.0	68.0	60.0	60.0
Gemma-2-9B	frac90	61.0	61.0	61.0	61.0	61.0	61.0	61.0	61.0	61.0	61.0
n=464	pr	61.0	61.0	61.0	61.0	61.0	61.0	61.0	61.0	61.0	61.0
Qwen2.5-7B	frac90	66.7	66.7	66.7	66.7	66.7	66.7	66.7	66.7	66.7	66.7
n=464	pr	74.1	74.1	74.1	74.1	74.1	74.1	74.1	74.1	74.1	74.1

below 0.5 at 96.0% depth against an onset at 60.0%, nine grid steps; on Gemma-2-9B at 97.6% against 61.0%, also nine. On Qwen2.5-7B the crossing is *undefined* rather than distant: the metric never falls below 0.5, its minimum over the sampled depths being 0.5937, so no such depth exists and the comparison is NaN. The sealed threshold sweep places its crossing at 96.3% under a 0.6 bar and 88.9% under 0.7. The sweep was sealed so that one threshold could not carry the verdict alone, and on that lens it is what makes the cell readable.

The reproducibility cost of this decomposition. The published description of signature (d) fixes the statistic but not the four manipulations, so every choice below had to be made by us. Table 14 lists them with the resolution taken, all of them fixed in writing before any of T1–T4 ran. Three are marked. M-1 is the choice that makes T3 a test of anything rather than a no-op; M-4 and M-5 are the tolerances the verdicts are read against, and a reader who disagrees with either can re-read the tables above under their own.

Table 14: Degrees of freedom in the decomposition, all resolved before the run. ⋆ marks the three that a verdict turns on: the object T3 randomises, and the two tolerances. M-9 and M-10 exist because T3 has two regimes in which it is provably a no-op.

	left underdetermined	resolved as
1⋆	Which Gram T3 randomises, $W_U W_U^\top$ or the centered M	The centered M, since $A_c = J_\ell^\top M J_\ell$ exactly and $W_U W_U^\top$ is not the object (d) depends on
2	How the random orthogonal Q is drawn	QR of a Gaussian with sign-fixed diagonal, Haar, reproducible from the seed alone
3	Effective dimensionality for a square J_ℓ	Rows as atoms, centered over rows, same code; an *analogue* and labelled so, with pr of the uncentered singular spectrum beside it
4⋆	Tolerance for "coincide"	One sampled grid index, not a fixed depth distance: the step is 4.00 on Gemma-2-2B but 2.44–4.88 on Gemma-2-9B and 3.70–7.41 on Qwen2.5-7B
5⋆	Criterion for "survives randomisation"	Median within one index step *and* seed spread within one index step; a third labelled outcome, unstable, for median-only
6	Seeds for Q	Ten, 0–9, with the spread reported and not the median alone (Table 13)
7	Threshold for "stops being near-identity"	0.5 on $\|J_\ell - I\|_F/\sqrt{d}$, the midpoint of the metric's own range, with a 0.3–0.7 sweep reported
8	What T1 should return	NaN: min–max normalising an exactly constant profile is undefined, and a non-constant T1 is a harness bug that blocks the rest
9	Whether T3's vacuity regimes gate its reading	Yes, as preconditions with numeric bars, evaluated before T3 is read
10	Which statistic adjudicates T3 if frac90 does not move	The participation ratio, *for that case only*. It did move, so the case never arose and the rule does not reach the disagreement we found
11	T4 coverage for Qwen2.5-7B, absent from the gate artifact	Recomputed from the lens tensor, a $d \times d$ Frobenius norm per layer, no model load
12	Which lenses	Three, for want of stored Jacobians, not for want of validation gates
13	Family count	Three families among our five fits, two among the runnable three

A cross-family test. If the unembedding contributes only its eigenvalue spectrum to signature (d) and not its alignment with J_ℓ, then differences in tokenizer and vocabulary between families are unlikely to be what produces the 50.1-point onset range, since those differences change M's orientation at least as much as its spectrum. That is a hypothesis and not a finding. The ladder cannot test it: its rungs vary everything at once, which is the confound Section 6 names, and the decomposition that could test it runs only where a stored Jacobian exists, which is three models from two families. T5 below is the one test of it we can run, and it is exploratory.

T5, the cross-family orientation swap, which is exploratory. T3 randomises M's eigenbasis with a Haar draw, which asks what *any* orientation does. The hypothesis above is about what *another family's* orientation does, and that is a manipulation we can perform directly, because Gemma-2-9B and Qwen2.5-7B have the same width ($d = 3584$) while differing in vocabulary by a factor of 1.68 (256,000 against 152,064) and in tokenizer, training and family. Writing $M_X = Q_X \Lambda_X Q_X^\top$, we form $M_{X \leftarrow Y} := Q_Y \Lambda_X Q_Y^\top$, which carries X's spectrum in Y's eigenbasis, and read signature (d) from $J_\ell^{(X)\top} M_{X \leftarrow Y} J_\ell^{(X)}$ with $J_\ell^{(X)}$ bit-identical to the fit. The spectrum is preserved to relative 7.60×10^{-16} and 4.12×10^{-16}, against a sealed tolerance of 10^{-10} that would have voided the arm.

Read through frac90, the statistic R-onset uses and therefore the one every onset in this paper is read through, the onset does not move in either direction: 61.0% on Gemma-2-9B under Qwen2.5-7B's orientation and 66.7% on Qwen2.5-7B under Gemma-2-9B's, both zero index steps from their own baselines. As in T3 the curve does move while the crossing does not, by 3.91×10^{-2} and 4.55×10^{-2} in frac90 against a quantisation of 2.79×10^{-4}, which is 140 and 163 quanta. Read through the participation ratio it moves in both directions, from 73.2% to 61.0% on Gemma-2-9B and from

77.8% to 74.1% on Qwen2.5-7B, three index steps and one. The same disagreement between the two dimensionality statistics that T3 found under a random orientation, within one model, appears here under a real orientation and across the family boundary: it is one phenomenon seen twice, not two findings. No sealed rule adjudicates between the statistics here either, and we supply none. Both facts hold: the paper's onsets are read through $\mathrm{frac90}$, which is the statistic that does not move under either manipulation, and through the participation ratio the onsets do move, by one to three grid steps depending on the lens.

So the hypothesis survives the one test we can run on it, under the statistic every onset in this paper is read through, and does not survive it under the other. It is one pair of models at matched width, it is exploratory, and no claim in the main text rests on it.

J ROBUSTNESS BATTERY

The battery summarized in Section 5.5 is reported here confound by confound, closing with a resolution limit that is not a confound at all but a property of the locating rule.

Since the onset rests on signature (d), what matters is whether (d) is stable at n=150, and having shown (b) to be fit-sensitive we measure rather than assume it. A checkpoint-preserving refit out to 464 prompts retains 13 lenses, its n=150 checkpoint bit-identical to the lens every Gemma-2-2B number traces to (maximum relative L_2 of 0 across the 25 source layers). Under R-onset at the sealed 0.25, the onset reads 60.0% at every checkpoint from n=20 to n=464: three times more fitting data than we used changes nothing. (d) settles later than its onset does, as it must given the 4.0-point quantisation, but it does settle: Pearson r against the n=464 profile reaches 0.999901 and the maximum c_{90} deviation falls to 0.0082 at n=75, past which the residual is at most 0.0052 and no longer decreasing monotonically. Across the threshold sweep the result is stable but not uniformly so: three of the seven thresholds are fixed by n=50; 0.20 and 0.30 settle only at n=100 and n=200, and 0.15 and 0.50 each still shift by one grid step as late as n=300.

Fit-size. The Gemma decomposition plateaus by $n \approx 35$ (e.g. variance-excess at n=35 vs. n=150: L8 3.298% → 3.169%, L20 2.711% → 2.608%); a matched-n control (n=126 vs. n=464, Gemma-2-9B) gives mid-band peak signature-(a) 0.1419 vs. 0.1367, so the level of signature (a) inside the band is not a fit-size artifact.

Convention. TransformerLens and HuggingFace activations agree to worst relative L2 3.258×10^{-7} and max $|\Delta|$ 1.831×10^{-4} (fp32), a PASS at tolerance 2×10^{-3}; we describe this as close numerical agreement, not bit-exactness.

DC contamination. Centering with $\mathrm{ln_final}$ moves the variance-excess minimum by 15.3 depth points (46.2% → 61.5%), and the DC component carries 0.223–0.509 of the variance across depth; the centered basis is the one used for every real-model occupancy and variance-excess number, while the planted-count gate decomposes raw synthetic activations (Appendix G.4).

Decomposition budget. As k increases 25 → 100, the gap between mid-depth variance-excess and the profile maximum at the same k *rises* (4.72 → 6.11 pp on Qwen2.5-7B and 7.56 → 8.03 pp on Gemma-2-9B), so the capacity profile is not an artifact of too small a decomposition budget.

Fit and eval regime. Refitting the lens on GSM8K (Cobbe et al., 2021) chain-of-thought (converged, $\mathrm{mean_rel_change}$ 0.240 → 0.036) leaves the Gemma-2-9B onset at 61.0%. A 2×2 fit×eval design gives mid-band peak signature-(a) of 0.0764/0.0384/0.1367/0.1198 (wt-fit×wiki / re-fit×wiki / wt-fit×GSM8K / re-fit×GSM8K): the elevation tracks the *evaluation* corpus rather than the fit corpus. One residual GSM8K-fit×GSM8K-eval interaction remains, which we report as unresolved and bounded.

Resampling the fit itself. The onsets above are point estimates from one averaged Jacobian each, which invites the question of how much of the cross-artifact spread is sampling noise in the fit. For two of our five lenses we can answer it directly, because they were fitted in disjoint prompt shards that were retained: Gemma-2-9B in 8 shards of 58 prompts and Qwen2.5-7B in 7 of 66–67. A shard is an exchangeable unit, so resampling shards with replacement resamples the fit. Per-prompt Jacobians were never retained, so a prompt-level bootstrap is not available and we do not construct one; Gemma-2-2B was fitted in a single process and has no shards at all, so it gets no interval here, only the convergence statement already reported.

Over 400 shard resamples, seed fixed in advance, the onset does not move at all: all 400 replicates return 61.0% on Gemma-2-9B and all 400 return 66.7% on Qwen2.5-7B, with no degenerate replicate in either. The 2.5–97.5 grid interval is therefore a single grid point on both. Read shard by shard, seven of Gemma-2-9B's eight individual shards return 61.0% and one returns 65.9%, one grid step away, while all seven of Qwen2.5-7B's return 66.7% exactly. A within-versus-between decomposition on c_{90} pools to a between-shard to within-bootstrap variance ratio of 15.28 on Gemma-2-9B and 7.52 on Qwen2.5-7B; the latter is close to the shard count, as a mean of seven exchangeable parts would give, and the former is inflated by that one outlying shard.

The consequence for Section 4 was checked in a form that could have cut the other way. The two intervals do not overlap, and resampling the prompts a lens was fitted on moves its onset by nothing on one model and at most one grid step in one shard of the other, against a 19.7-point spread across our five fits and 50.1 across the ladder. The spread between artifacts is not fit-sampling noise. It remains an interval on two lenses of five, at the shard level and not the prompt level, and it says nothing about the published comparison lenses, whose shards we do not have.

Depth resolution, and why GPT-2 small is reported as an interval. R-onset returns the first depth at which a min–max normalized curve crosses a threshold, so the precision of the answer cannot exceed the spacing of the depth grid, and that spacing is set by the model's layer count. Every rung admitted by the screen of Section 4 is measured at 25 depths; because the subsample is spread over the model's own layer count, the nominal step is 4.0 points, but samples land on integer layers, so the actual step varies about it and reaches 5.1 on some admitted rungs, and "25 depths" bounds the grid rather than fixing it. GPT-2 small has only 11, a 9.1-point grid, and is the case that motivates the screen: at the sealed threshold it returns 45.5%, but across the grid $\{0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.50\}$ it moves from 18.2% to 72.7%, a 54.5-point excursion. The relevant comparison is against the whole screened ladder rather than a handful of models: across the 22 model–lens artifacts the same sweep has a median width of 22.2 points (quartiles 16.0 and 31.5), running from 7.9 on OLMo-3-32B to 68.0 on Gemma-3-1B, with 13 of the 22 at or below 24.0 and exactly one wider than GPT-2's 54.5. Per-rung widths are the *sweep* column of Table 3.

The cross-family result itself does not rest on the threshold. Re-running R-onset at each grid value in turn, the onset span across the 22 model–lens artifacts is 69.8, 57.0, 45.0, 50.1, 61.6, 56.8 and 61.7 points at 0.10, 0.15, 0.20, 0.25, 0.30, 0.35 and 0.50: never narrower than 45.0 points, so no choice on the grid collapses the spread of Section 4, and the 0.25 we fixed in advance is neither the widest nor the narrowest reading available.

Two things follow, and a third qualifies them. First, GPT-2's onset is weakly determined and we report the interval rather than the value; a single number from an 11-depth profile would overstate what the rule can resolve. Second, a coarse grid is *sufficient* to produce a wide sweep, so any onset comparison across models of different depths is partly a comparison of resolutions, and onsets from short models should be reported as intervals.

Third, a coarse grid is not *necessary* for a wide sweep, so sweep width cannot be read back as evidence of poor resolution. Gemma-3-1B is measured at the full 25 depths and still moves 68.0 points across the same threshold grid, wider than GPT-2's 54.5 on less than half the step size. Whatever produces that width in Gemma-3 is not the grid (Appendix C). The screen of Section 4 is therefore a resolution criterion only, and it should not be mistaken for a filter on how well determined a rung's onset is: rungs can pass it and still carry wide intervals, which is why the sweep column is reported for every rung rather than only for the excluded ones.

K A CONTINUOUS LOCATOR, PROPOSED ALONGSIDE R-ONSET

R-onset returns the first *sampled* depth at which a min–max-normalized curve crosses 0.25, so its resolution can never beat the depth grid. That is the limitation Appendix J documents. We fit a four-parameter logistic $g(x) = a + (b - a)/\big(1 + e^{-(x-x_0)/s}\big)$ to the same normalized curve by least squares and read the crossing off the fitted function, at the depth where the fitted curve first reaches 0.25 of its own fitted range. That makes it the interpolated analogue of R-onset rather than a different target. Its plan was sealed in advance and it is reported as a proposal: every onset in this paper stays R-onset on signature (d), and no number we report is re-derived from this.

Validated on curves whose transition is known. On 200 synthetic logistic curves with x_0 drawn across the interior of the depth range, sampled on the same 25-point grid and carrying Gaussian noise at the residual scatter of the real fits (0.0321), the estimator converged on 200 of 200. Its median absolute error is 0.479 depth points against a one-grid-step bar of 4.17, and the empirical coverage of its 95% interval is 0.915 against a sealed band of $[0.90, 0.99]$. Both bars are met, so we report it as usable.

What it says about our own rungs. It converges on all 31 stored signature-(d) profiles and publishes an interval for each. The continuous onset runs systematically *earlier* than R-onset, by a median of 2.61 depth points (quartiles 0.58 and 3.91), which is exactly what a first-sampled-depth rule does: it rounds the crossing up to the next grid point. On the three gated fits the two agree closely and the intervals are tight: Gemma-2-2B 60.0% against 58.97% $[57.8, 60.2]$, Gemma-2-9B 61.0% against 59.27% $[57.7, 60.8]$, and Qwen2.5-7B 66.7% against 63.05% $[61.5, 64.6]$, with standard errors of 0.60, 0.78 and 0.81 depth points.

Two diagnostics that were built independently partly agree about which rungs are weak. Gemma-3-27B carries a standard error of 24.98 depth points, GPT-2 16.77, Gemma-3-12B 16.49 and Gemma-3-4B 11.65, against 0.60 to 0.81 on the three gated fits, and the three Gemma-3 rungs among them also carry wide threshold sweeps. The agreement is not complete: Gemma-3-1B, the widest sweep in the ladder at 68.0 points, has a standard error of only 2.31. Sweep width and fit uncertainty are computed from different things, one a threshold excursion and the other a residual scatter, and they overlap on the weakest rungs without identifying the same set. The fitted intervals also support the choice in Appendix J to report GPT-2 as an interval rather than a value: its fitted interval is the second widest in the pool.

What this does not license. A tighter locator does not make a cross-family onset portable. It sharpens each rung and leaves the spread between rungs where it was, and the disagreement between rules on one model (Table 4) is untouched by it, because that disagreement is between signatures rather than between crossings of one curve.

L A J-SPACE PATCHING PROBE ON GEMMA-2-2B

A steering pilot sealed earlier, which moved Gemma-2-2B along its top J-lens directions on 64 prompts, did not resolve: its sanity gate (top directions outweighing bottom ones at more than 80% of sampled layers) was met at only 20–76% across four settings, and we draw nothing from it. We therefore ran a second causal probe under a plan sealed before it ran. Its question is where the J-space component of the residual state becomes sufficient to carry a verbalized answer. Prompts are "The capital of X is" for 69 countries; the 32 that Gemma-2-2B answers correctly at top-1 form 96 source–target pairs with different answers. At each of the lens's 25 source layers, at the final position, we replace a component of the target's state with the source's and record the fraction of pairs whose top-1 token becomes the source's answer. The J-space component $c(h)$ is the reconstruction of h by non-negative gradient pursuit with k=16 atoms, the original account's k for a J-space component, over the unit rows of $W_U^\top J_\ell$ (uncentered, our n=150 lens). The conditions are J, $h_t - c(h_t) + c(h_s)$; C, the complement swapped instead, $c(h_t) + h_s - c(h_s)$; R, as J with a random unit dictionary of the same size; F, the whole state; and a no-op.

Sanity gates. Both pass: the whole-state swap at the last source layer transfers 0.969 of pairs against a bar of 0.90, and the no-op leaves every target's answer in place at every layer.

Result. Read with the rule R-onset uses, the J-space swap's transfer rate has its onset at 80.0% depth, 20 points after the signature-(d) onset of 60.0%; it reaches 0.75 of pairs at 88.0% (Table 15). The random-dictionary and complement swaps transfer nothing at any depth, although the J-space component holds only 0.03–0.12 of the state's energy, so what transfers is specific to the J-space component rather than any component of that size. Part of the gap belongs to the site: at the final position even the whole state has its onset at 72.0%, since before then the answer is carried at earlier positions that this probe leaves untouched. The finding is therefore where the J-space component becomes sufficient for the verbalized answer at one site, on one model and one relation: late, and later than the spectral onset.

Table 15: The J-space patching probe, Gemma-2-2B, 96 pairs: fraction of pairs whose top-1 token becomes the source's answer after each swap at the final position. Every condition is 0.000 at every depth below 64%. The last column is the median share of the state's energy in its J-space component.

depth	J	C	R	F	$\|c\|^2/\|h\|^2$
64.0%	0.000	0.000	0.000	0.000	0.046
68.0%	0.000	0.000	0.000	0.000	0.053
72.0%	0.000	0.000	0.000	0.406	0.070
76.0%	0.052	0.000	0.000	0.448	0.082
80.0%	0.240	0.000	0.000	0.833	0.103
84.0%	0.490	0.000	0.000	0.812	0.119
88.0%	0.750	0.000	0.000	1.000	0.114
92.0%	0.635	0.000	0.000	0.969	0.109
96.0%	0.542	0.000	0.000	0.969	0.106