THE DIMENSIONAL COST OF CONTEXT IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-rank structure in language-model activations is often taken as evidence that the residual stream can be narrowed. However, a subspace that works for one prediction may not work for another. We define the context requirement dimension as the smallest rank of a (linear) map that preserves the model's output distributions within a fixed tolerance. We show that a prediction may depend on few directions, but these directions vary across contexts and together cover most of the residual width. At the output, preserving the distributions within our tolerance requires roughly 67–83% of the residual width across 21 models from nine families. The requirement is lower inside the network but generally increases towards the output. During generation, narrowing the stream shifts output towards frequent tokens. Therefore, local low dimensionality gives a misleading estimate of the width needed across a corpus.

1 INTRODUCTION

Several measurements make language-model states appear low-dimensional. Activation variance can concentrate in a few directions (Timkey & van Schijndel, 2021; Sun et al., 2024; Salfati, 2026). Estimates of intrinsic dimension can also be much smaller than the residual width (Valeriani et al., 2023; Cheng et al., 2023). Around a single prediction, the output Fisher metric may have low effective rank (Park et al., 2026; Wang et al., 2026). Reducing residual width could lower computational cost, but the retained space must preserve predictions across prefixes (Ashkboos et al., 2024; Sakr & Khailany, 2024).

These measurements do not test whether the same directions work across contexts. One prediction may depend on a small set of directions and another prediction on a different set. Each prediction can therefore appear low-dimensional while preserving predictions across a corpus requires most of the residual width. Corpus-level Jacobian analyses already show that geometry aggregated over many inputs can be broad (Sarkar & Deka, 2026; Gurnee et al., 2026). They do not measure the rank required by one reconstruction to preserve full output distributions on unseen prefixes.

In this paper, we study how many directions of the residual stream are required, both inside the network and at the final residual state, to preserve the model's next-token distributions. We also ask why this requirement can be much larger than the dimension measured around one prediction.

We answer these questions by fitting a rank-k reconstruction at a chosen depth, leaving all model weights unchanged. We evaluate the same map on unseen prefixes using KL between the intact and reconstructed output distributions. We call the smallest tested rank meeting a fixed mean-KL tolerance the *context requirement dimension*.

Our contributions are:

- Across 21 models from nine families, we show that the smallest passing rank at the output is 66.7–83.4% of model width. Inside the network, the requirement is lower but remains substantial and generally increases towards the output.
- We trace this high shared rank to prefix-specific sensitivity. Each prefix is sensitive to few directions, but those directions vary across prefixes, even with the same next token. We prove that narrow local sensitivities can collectively span the full width. The prediction-relevant geometry is therefore not one narrow global subspace, but many narrow prefix-specific subspaces with a broad combined span.

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

- We find the same local-to-global gap with PCA, output-Fisher-weighted PCA, and KL-trained reconstructions. This limits fixed-basis compression and hidden-state distillation as one narrow global basis does not preserve the directions needed across prefixes.
- We show that low-rank reconstruction of final states shifts generation towards frequent tokens. At half width, the median share outside the 1,000 most frequent token types falls by 9.3 percentage points.

The rest of the paper is organized as follows. Section 2 reviews related work, Section 3 measures shared width at the output and across depth, Section 4 studies prefix-specific sensitivity and Section 5 evaluates generation. We discuss implications and limitations in Section 6.

2 Related Work

Local and shared dimensions. Low dimension can refer to different objects. Intrinsic-dimension estimates describe activation geometry and how it changes across layers (Valeriani et al., 2023; Cheng et al., 2023; 2025; Schulte & Rügamer, 2026). Output-aware methods instead measure which directions affect one prediction through the unembedding or local softmax Fisher metric (Park et al., 2024; 2026; Wang et al., 2026). Sarkar & Deka (2026) measure the effective rank of an expected Jacobian outer product across inputs. Gurnee et al. (2026) construct a representation space from Jacobians averaged across prompts and conclude that few of its directions are active at one position, although the full space spans much of the residual stream at later layers. Prior work therefore characterizes either local geometry or geometry aggregated across inputs. Instead, we measure a different quantity: the rank required by one reconstruction to preserve full output distributions on unseen prefixes, both inside the network and at the output.

Fixed-subspace compression. This distinction matters the most for compression. For instance, activation variance does not need to match predictive importance (Timkey & van Schijndel, 2021; Salfati, 2026; Shihab et al., 2026), and as such, can completely destroy next-token predictions. Existing width-reduction methods can compensate through learned rotations, calibration, or retraining (Timkey & van Schijndel, 2021; Salfati, 2026; Shihab et al., 2026). In this paper, instead, we keep the model fixed, insert one shared reconstruction at a chosen depth, and measure its effect on the output distribution after the remaining blocks have run.

Token frequency under compression. Compression may also affect some outputs more than others. Previous work finds larger errors on infrequent examples (Hooker et al., 2019; Ahia et al., 2021). Language models can encode token frequency in an output bias or in the final state, while other directions counter this preference when the prefix supports another token (Meister et al., 2023; Macocco et al., 2025; Liu et al., 2026). We leverage these findings and test whether low-rank reconstruction similarly shifts generation towards frequent tokens.

3 Width Required Across Contexts

In this section, we measure how many directions a reconstruction shared across prefixes must retain to preserve next-token predictions. We first apply the reconstruction to the final output state, then repeat the measurement at intermediate layers.

3.1 Measurement

Let $c = (w_1, \ldots, w_t)$ denote the token prefix used to predict w_{t+1}. Let $h(c) \in \mathbb{R}^D$ be the final state after normalization, immediately before the output head, where D is the model width. We replace this state with $R_k(h(c))$, while keeping the model and output head fixed. For a given rank k, the reconstruction is

$$R_k(h) = \mu + A_k(h - \mu), \qquad \operatorname{rank}(A_k) \le k,$$

where μ is the mean final state on the fitting data. The reconstruction preserves the fitting mean μ and retains at most k directions of variation around it. Let W_U be the unembedding matrix, and $b \in \mathbb{R}^{|\mathcal{V}|}$ the output bias. The reconstructed prediction is

$$p_k(\cdot \mid c) = \operatorname{softmax}(W_U R_k(h(c)) + b)\,.$$

Let $p_\theta(\cdot \mid c)$ be the intact model prediction. We compare it with $p_k(\cdot \mid c)$ using Kullback-Leibler (KL) divergence:

$$\Delta_k(c) = D_{\mathrm{KL}}(p_\theta(\cdot \mid c) \,\|\, p_k(\cdot \mid c)) .$$

Let $\mathcal{K} \subseteq \{0, \ldots, D\}$ be the set of evaluated ranks. For a set of prefixes $\mathcal{C}$, the *context requirement dimension* is the smallest rank in $\mathcal{K}$ whose average distortion is below a tolerance τ:

$$k_\tau(\mathcal{C}) = \min \left\{ k \in \mathcal{K} : \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \Delta_k(c) \leq \tau \right\} .$$

In our experiments, we use $\tau = \log(1.05) = 0.0488$ nats. This corresponds to a 5% increase in perplexity when next tokens are distributed according to the intact model.

For each rank k, the same reconstruction R_k is applied to every prefix in $\mathcal{C}$. Thus, $k_\tau(\mathcal{C})$ measures the rank required by one shared reconstruction, not by a separate subspace for each prediction. This tests whether the final states admit a common low-dimensional representation.

At each rank k, we construct A_k in four ways: PCA, output-Fisher-weighted PCA, paired-Fisher training, and KL training. All four methods are evaluated with the same KL distortion $\Delta_k(c)$ defined above. The Fisher-based methods estimate how strongly a state error changes the next-token distribution.

- **PCA** retains the directions with the largest activation variance. It uses $A_k = V_k V_k^\top$, where the columns of V_k are the k principal directions of the centred states. It does not use the output head or the model's predictions.
- **Output-Fisher-weighted PCA** weights reconstruction errors by their estimated effect on the output distribution before applying PCA. It uses $A_k = S^{-1} U_k U_k^\top S$, where S is the square root of the regularised corpus-average Fisher matrix and the columns of U_k are the principal directions of $S(h - \mu)$.
- **Paired-Fisher training** learns the reconstruction using a quadratic approximation to KL, with a separate Fisher matrix for each prefix. It uses $A_k = UV^\top$, with $U, V \in \mathbb{R}^{D \times k}$.
- **KL training** learns the reconstruction by directly minimising KL between the intact and reconstructed predictions. It also uses $A_k = UV^\top$, with learned matrices $U, V \in \mathbb{R}^{D \times k}$.

We use KL training for the main measurement because it minimises the distortion used to define $k_\tau(\mathcal{C})$. The other three methods test whether the result depends on how A_k is constructed. PCA mostly test whether variance is a good proxy for predictive importance.

The Fisher-based methods use the softmax geometry studied by Park et al. (2026). Output-Fisher-weighted PCA adapts metric-weighted low-rank approximation to residual states (Hsu et al., 2022). Appendix A defines all four methods and gives the fitting procedure.

3.2 Final-State Requirement

We test 21 models from nine families on WikiText-103. We fit the reconstructions on 300 articles, select trained maps using 75 validation articles, and evaluate them on 75 separate test articles. We report rank as a fraction of model width, k/D. Appendix C describes how we prepare the data and how we fit the maps.

Figure 1 shows first passing ranks between 66.7% and 83.4% of model width. Qwen3.5-2B and Qwen3.5-4B pass at two-thirds width while the other 19 models first pass between 70% and 83.4%. Appendix D supports this conclusion across data samples, model depths, training stages, and text distributions.

Table 1 compares the four fitting methods. KL training gives the lowest or a tied first passing rank in most models. Output-Fisher-weighted PCA and paired-Fisher training often give similar ranks, while ordinary PCA usually requires more width. The fitting method changes the threshold, but high required ranks occur with all four methods.

More importantly, larger models do not consistently pass at smaller fractions of their width.

Of course, these are the best maps found by our fitting procedure, not lower bounds on all possible reconstructions. Better-fitted linear maps or nonlinear reconstructions may require fewer dimen-

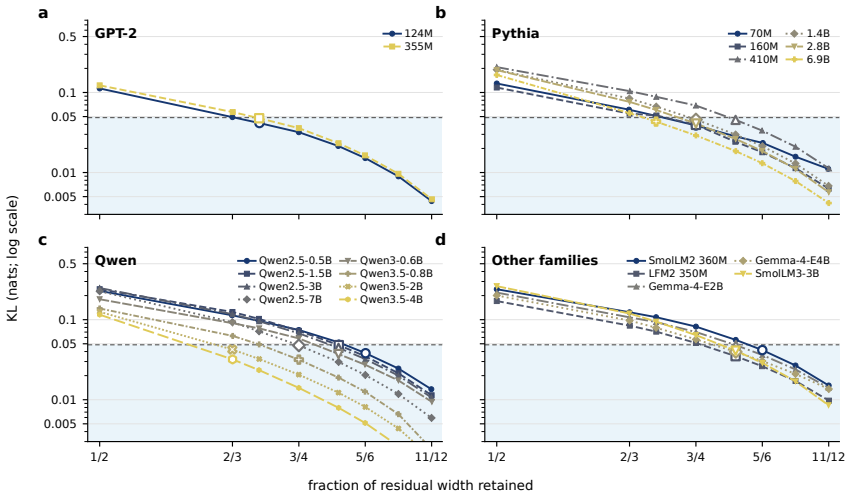

Figure 1: Mean KL on unseen WikiText-103 prefixes as the retained width k/D increases. Each curve represents one model. The shaded region marks the 5% perplexity tolerance.

Table 1: Smallest retained-width fraction k/D meeting the 5% perplexity tolerance. Lower is better.

Model	PCA	Fisher PCA	Paired Fisher	KL training
GPT-2	.80	.75	**.70**	**.70**
GPT-2 Medium	.80	.75	.75	**.70**
Pythia-70M	$5/6$	$23/24$	$23/24$	**.75**
Pythia-160M	.80	.75	**.70**	.75
Pythia-410M	$7/8$	**.80**	**.80**	**.80**
Pythia-1.4B	$5/6$	**.75**	**.75**	**.75**
Pythia-2.8B	.80	**.75**	**.75**	**.75**
Pythia-6.9B	.75	**.70**	**.70**	**.70**
SmolLM2-360M	$7/8$	$\mathbf{5/6}$	$\mathbf{5/6}$	$\mathbf{5/6}$
SmolLM3-3B	$7/8$	**.80**	**.80**	**.80**
LFM2-350M	$5/6$	**.80**	**.80**	**.80**

Model	PCA	Fisher PCA	Paired Fisher	KL training
Qwen2.5-0.5B	$7/8$	$\mathbf{5/6}$	$\mathbf{5/6}$	$\mathbf{5/6}$
Qwen2.5-1.5B	$7/8$	**.80**	**.80**	**.80**
Qwen2.5-3B	$7/8$	**.80**	**.80**	**.80**
Qwen2.5-7B	.80	**.75**	.80	**.75**
Qwen3-0.6B	$5/6$	**.80**	**.80**	**.80**
Qwen3.5-0.8B	.80	**.75**	**.75**	**.75**
Qwen3.5-2B	.75	$\mathbf{2/3}$	$\mathbf{2/3}$	$\mathbf{2/3}$
Qwen3.5-4B	.70	$\mathbf{2/3}$	$\mathbf{2/3}$	$\mathbf{2/3}$
Gemma-4-E2B	**.80**	$5/6$	**.80**	**.80**
Gemma-4-E4B	**.80**	**.80**	**.80**	**.80**

sions. However, the results are constant enough accross size and families to suggest that the high shared rank is a general property of the final states, not an artifact of one fitting method.

The required rank grows during pretraining. Across three Pythia models, it rises from below 10% of width to 75–80%, largely after most of the loss reduction has occurred. The full training trajectories examined here cover Pythia only.

The requirement also depends on the text distribution. Pile code requires the least width and the mixed Pile the most. At five-sixths width, 43 of 54 cross-corpus applications exceed the KL tolerance. A reconstruction fitted on one corpus therefore does not reliably preserve predictions on another. We hypothesise that vocabulary and regularity contribute to these differences, but the corpus comparison does not separate their effects. Appendix D gives the training curves and cross-corpus results.

Finally, prefixes followed by infrequent token types have 3.5 to 5 times higher KL than those followed by a top-100 type. However, the top-100 group still usually requires 66.7–70% of the width. Therefore, target-token frequency does not by itself explain the high shared rank (Appendix Table 6).

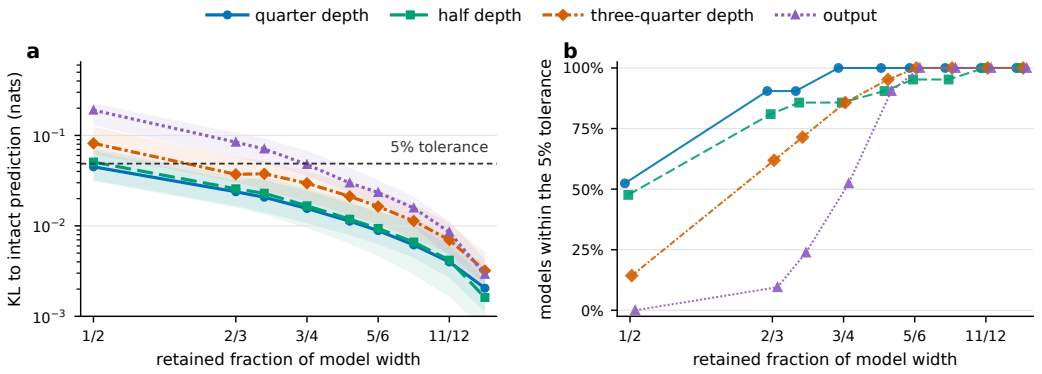

Figure 2: Requirement across depth. **(a)** KL against the retained fraction of model width. Lines show medians across 16 models and bands show interquartile ranges. **(b)** Fraction of models below the 5% perplexity tolerance. Intermediate states tolerate lower ranks than output states, but the advantage decreases towards the output.

3.3 Requirement Across Depth

In the previous section, we measured the context requirement dimension at the final residual state. We now repeat the measurement at intermediate layers to see how the requirement changes across depth. We repeat the measurement at one-quarter, one-half and three-quarters of the network depth. At each depth, one shared reconstruction is fitted at the input of a block. The remaining blocks and the output head then run unchanged, and we measure KL on the same unseen articles. As reference, previous work reports two different depth profiles. Causal dimensionality remains approximately constant across the layers tested by Sarkar & Deka (2026), whereas the effective dimension of the Jacobian-defined J-space increases through the middle layers and approaches the residual width near the output (Gurnee et al., 2026).

Figure 2 shows the KL across retained widths and the share of models that pass the tolerance. At the same width, reconstructing an intermediate state destroys less the final prediction than reconstructing the final state. This difference narrows with depth, so later layers generally require more width. This is consistent with the intuition that the remaining blocks can recover from some information removed earlier in the network and the findings of Gurnee et al. (2026). However, the advantage of intermediate states decreases towards the output. At three-quarter depth, 13 of the 21 models in Figure 2 pass at two-thirds width, and all pass by five-sixths.

4 Context-Specific Sensitivity

Section 3 found that our shared linear reconstructions retain much of the model width to preserve predictions. Previous work finds that, for one prefix, the output Fisher metric is often concentrated in a few residual-stream directions (Park et al., 2026; Wang et al., 2026). These two findings are compatible if the sensitive directions change between prefixes. This section studies whether different prefixes rely on different sensitive directions, even when their observed next token is the same.

4.1 Local and Shared Sensitivity

Does low-dimensional sensitivity at each prefix imply that the same low-dimensional subspace works across prefixes? Not necessarily. Each prefix can be sensitive to few directions without those directions being shared across prefixes.

For prefix c_i, let $p_i = p_\theta(\cdot \mid c_i)$ be the intact next-token distribution. Its local output Fisher metric is

$$G(c_i) = W_U^\top [\mathrm{diag}(p_i) - p_i p_i^\top] W_U,$$

which we abbreviate as $G_i = G(c_i)$. A direction has a large value under metric G_i when a small change along that direction strongly affects the output distribution. For positive weights π_i summing

to one, the corpus-average metric is

$$\overline{G} = \sum_i \pi_i G_i.$$

If the dominant directions are similar across prefixes, $\overline{G}$ remains concentrated. If they differ, $\overline{G}$ can be broad even when every G_i is concentrated. We measure this concentration by the participation rank defines for a positive semidefinite matrix M as

$$\mathrm{PR}(M) = \frac{\mathrm{tr}(M)^2}{\mathrm{tr}(M^2)}.$$

Proposition 1 (Local-to-global Fisher separation)**.** *For every width $D \geq 1$ and every $\eta > 0$, there is a biasless, full-support softmax model with D prefixes such that every local Fisher matrix has exact rank D while*

$$\mathrm{PR}(G_i) < 1 + \eta \quad \textit{for every } i, \qquad \mathrm{PR}\left(\frac{1}{D}\sum_{i=1}^{D} G_i\right) > D - \eta.$$

The proof is provided in Appendix B. The proposition gives an exact example of the local-to-global broadening studied empirically through expected Jacobian rank (Sarkar & Deka, 2026). The proposition shows that sensitivity can concentrate along one direction at each prefix while the average spreads across all D directions. The construction does not depend on the observed next-token labels, which may all be identical. We next test whether sensitive directions vary in trained models and whether this variation matters for reconstruction.

4.2 Trained Models

We test two aspects of the local prefix sensitivity. First, we measure how many directions capture the global sensitivity as more prefixes are combined. Second, we test whether the state component removed by R_k has a larger effect under the Fisher metric of its own prefix or, instead, is equally important for another prefix with similar entropy and target-token frequency.

Accumulation across prefixes. We first ask whether different prefixes are sensitive to the same residual-stream directions. For the experiments, we report how many directions capture 90% of the average Fisher sensitivity. Unlike participation rank, this gives a direct count of directions. For a set of prefixes $\mathcal{C}$, the average Fisher matrix is $\overline{G}_{\mathcal{C}} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} G(c)$. Let $\lambda_1(\overline{G}_{\mathcal{C}}) \geq \cdots \geq \lambda_D(\overline{G}_{\mathcal{C}})$ be its eigenvalues. We define

$$r_{90}(\mathcal{C}) = \min\left\{k : \frac{\sum_{j=1}^{k}\lambda_j(\overline{G}_{\mathcal{C}})}{\mathrm{tr}(\overline{G}_{\mathcal{C}})} \geq 0.9\right\}.$$

Thus, $r_{90}(\mathcal{C})$ is the number of directions containing 90% of the output sensitivity across the prefixes in $\mathcal{C}$. This measures Fisher sensitivity, not the rank needed to meet the KL tolerance in Section 3.

We measure $r_{90}(\mathcal{C})$ in 21 models. We compare one prefix, groups of prefixes followed by 1–64 different token types, and a sample of 4,096 corpus prefixes. In the comparison across token types, the number of prefixes is fixed. For each group, we average its Fisher matrices and measure r_{90}. The results are reported in Figure 3a.

Across the 21 models, the median r_{90} rises from 1.5% of model width for one prefix to 18% for prefixes followed by the same token, and to 55.7% for the corpus sample. For one prefix, the local Fisher participation rank ranges from 4 to 13 across models. Sensitive directions therefore vary even between occurrences of the same next token which is not surprising given the fact the state trajectories through depth do not have to be the same or follow the same *path* to the same next token type.

We also compare the subspaces formed by the eight leading Fisher directions. Their mean overlap score is 0.34 for prefixes followed by the same token and 0.12 for different tokens in the same frequency class, where 1 means identical subspaces. Appendix E reports these overlap and grouping controls.

Token identity affects significatively the directions which is actually a more surprising result, but does not determine them.

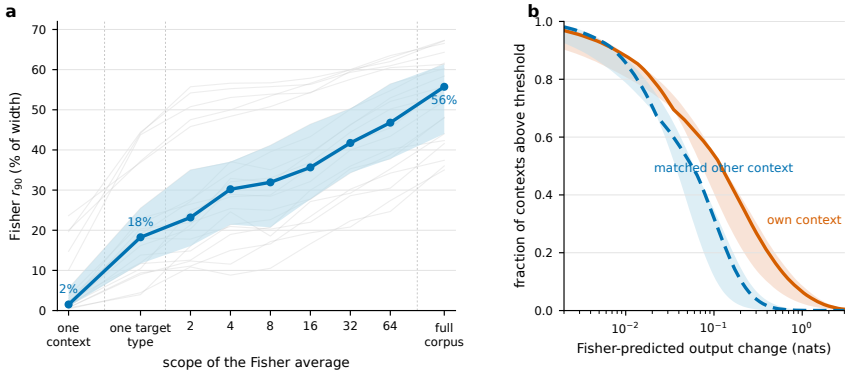

Figure 3: Prefix-specific output sensitivity. **(a)** The number of directions needed to capture 90% of output sensitivity increases from one prefix to the full corpus. **(b)** A state component removed for one prefix has a larger estimated effect under that prefix's Fisher metric than under the metric of a matched prefix.

Prefix-specific reconstruction errors. Do the components removed by reconstruction matter particularly to their own prefix? We test this using output-Fisher-weighted PCA at rank $D/2$. For each prefix c, $e(c)$ is the vector removed from its final state:

$$e(c) = h(c) - R_{\frac{D}{2}}(h(c)).$$

We estimate the resulting KL using the prefix's Fisher matrix:

$$\widetilde{\Delta}(c) = \frac{1}{2} e(c)^\top G(c) e(c).$$

We test this on every held-out position which is about 104,000 per model. For each position, we first score the removed vector with the Fisher matrix of its own prefix. We then score the same vector three more times, each time using the Fisher matrix of a randomly reassigned prefix from the same entropy quintile and target-frequency class. The removed vector remains unchanged. With this protocol, if prefix-specific sensitivity did not explain the effect of the reconstruction, reassignment would have little effect.

Figure 3b reports the results for all test positions, with the three matched reassignments averaged for each position. Using another prefix's Fisher matrix lowers the mean estimated KL by 69–70% across GPT-2, Pythia-410M, and Qwen2.5-0.5B. The same removed vector therefore has a larger estimated effect under its own prefix's sensitivity. Prefixes with larger Fisher estimates also tend to have larger measured KL, with Spearman correlations of 0.94–0.998.

The shared reconstruction therefore removes directions that matter specifically to the prefix being processed. This connects the two earlier findings: one shared reconstruction needs many directions because different prefixes are harmed by removing different directions. Model-level values and controls are reported in Appendix E.

In conclusion, individual predictions have concentrated local sensitivity, but the sensitive directions vary between prefixes. This variation occurs within one target-token type and determines which reconstruction errors damage the output. A shared bottleneck needs many directions because it must preserve predictions under many different local Fisher geometries. These experiments analyze the final state and whether the same mechanism explains the full depth profile remains open.

5 Low-Rank Reconstruction During Generation

Sections 3 and 4 evaluated next-token predictions on held-out prefixes. During generation, each sampled token becomes part of the next prefix, so each prediction affects those that follow. We ask whether applying the shared low-rank reconstruction to the final output state changes which tokens the model generates. Language models can encode token frequency in an output bias or in the

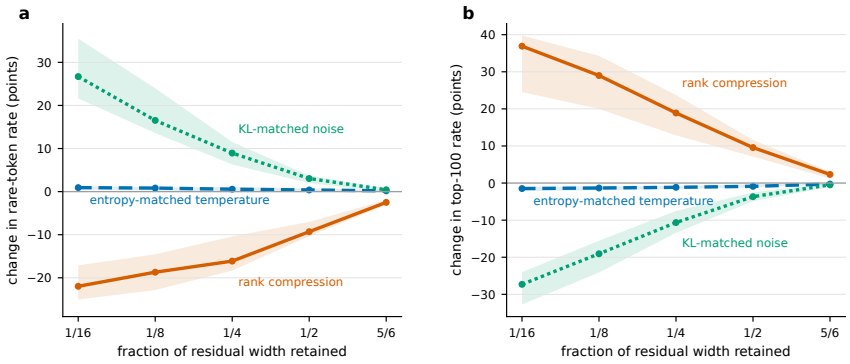

Figure 4: Change in generated-token frequency relative to the intact model. **(a)** Share outside the 1,000 most frequent types. **(b)** Share among the 100 most frequent types. Lines show cross-model medians over 21 models and bands show interquartile ranges. Low-rank reconstruction is compared with entropy-matched temperature and KL-matched isotropic noise.

final state. Other directions adjust predictions to the current prefix (Meister et al., 2023; Liu et al., 2026; Macocco et al., 2025). We hypothesise that removing some of these directions weakens this adjustment, increasing the share of frequent tokens.

5.1 Protocol

We test all 21 models, ranging from 70M to 7B parameters. For each model, we generate 128 tokens from the same 200 held-out WikiText-103 prefixes using nucleus sampling ($p = 0.9$, temperature 1) and three random seeds. Before generation, we fit one output-Fisher-weighted PCA basis as described in Section 3 and construct reconstructions at five ranks. We run generation separately at each rank, applying the same reconstruction to the post-normalization output state at every step. The state changes with the prefix, but the reconstruction does not.

We rank token types by their frequency in WikiText-103 validation text. We then measure the share of generated tokens inside the top 100 and outside the top 1,000. We use two controls. The temperature control adjusts the intact model's temperature to match the reconstruction's average output entropy, a measure of uncertainty. The noise control adds random noise with equal variance in every final-state direction, adjusting its scale to match the reconstruction's average KL from the intact model. Both controls are calibrated on held-out text before generation, and their settings remain fixed during generation. These controls test whether the frequency shift comes from confidence or from a generic disturbance rather than low-rank reconstruction. Appendix F provides the full protocol.

5.2 Frequency Shift

Figure 4 compares the three interventions across retained widths.

As retained width decreases from $5D/6$ to $D/16$, the median reduction in the outside-top-1,000 share grows from 2.5 to 22.0 percentage points relative to intact generation. The median increase in the top-100 share grows from 2.3 to 36.9 points. Neither control produces this pattern. Temperature scaling barely changes either token-frequency share, while KL-matched noise shifts generation towards infrequent tokens. Matching average entropy or KL on held-out prefixes therefore does not reproduce the frequency shift caused by low-rank reconstruction. Appendix Table 7 summarizes the changes at half width.

Previous studies found larger errors on long-tail images and infrequent machine-translation sentences (Hooker et al., 2019; Ahia et al., 2021). Our experiment shows a related bias during generation: low-rank reconstruction favours frequent tokens, whereas noise calibrated to the same average KL favours infrequent tokens. Average KL on held-out prefixes therefore does not determine the direction of the frequency shift during generation.

6 Discussion and Limitations

Prior work finds narrow geometry around individual predictions and broader Jacobian-based geometry across inputs (Park et al., 2026; Sarkar & Deka, 2026; Gurnee et al., 2026). We test whether this difference matters when reconstructing states. Across 21 models, our fitted maps first preserve output distributions within tolerance at 67–83% of final-state width. Different prefixes rely on different directions, even when they are followed by the same token. One shared reconstruction has to preserve predictions for all of them. Intermediate states generally tolerate lower ranks, although the required fraction of width remains substantial.

Low-rank reconstruction also shifts generation towards frequent tokens, extending earlier observations of disproportionate compression errors on infrequent examples (Hooker et al., 2019; Ahia et al., 2021). Interventions calibrated to the same average KL on held-out text can produce opposite frequency shifts during generation. Average KL alone therefore does not describe how vocabulary use changes.

Therefore, low local dimension is insufficient evidence for choosing a narrow shared projection. This matters when compression or hidden-state distillation uses one projected teacher space across prefixes (Sun et al., 2019).

However, our fitted ranks are not global lower bounds: better-fitted maps, nonlinear maps, or prefix-selected maps may need less width.

Few sensitive directions at each prefix do not imply that one small set of directions will preserve predictions across prefixes.

AI usage statement

In this work, we have used AI tools all along the research process. The original idea and the research question were formulated by the authors, and submitted to a workshop where the paper has been rejected. We used LLM to recenter the research question thanks to the feedback received from the workshop, and for an additional literature reviews as the core topic of the paper evolved rapidely in 2026. We did not use LLM to establish the experimental plan, but we used LLM to verify and consolidate the coherence of the experimental plan. We used LLM to write the code for the new experiments and the generation of the figures. The experimental code has be audited both by other AI tools and by the authors. Regarding writing, we used LLM to help with reformulation, type checking, and grammar checking. Regarding mathematics, we used LLM to figure out the most common notations and formalism in the literature. The proof of Proposition 1 is from the authors but we used LLM to verify the coherence and validity of the proof.

References

Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. The low-resource double bind: An empirical study of pruning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3316–3333. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-emnlp.282. URL https://aclanthology.org/2021.findings-emnlp.282/.

Saleh Ashkboos, Maximilian Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. SliceGPT: Compress large language models by deleting rows and columns. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/316648eb8b4ffb6010f531b07848c300-Abstract-Conference.html.

Emily Cheng, Corentin Kervadec, and Marco Baroni. Bridging information-theoretic and geometric compression in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12397–12420, 2023. URL https://aclanthology.org/2023.emnlp-main.762/.

Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Lei Yu, Alessandro Laio, and Marco Baroni. Emergence of a high-dimensional abstraction phase in language

transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/57568e093cbe0a222de0334b36e83cf5-Abstract-Conference.html.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, 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*, 2026. URL https://transformer-circuits.pub/2026/workspace/index.html.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019. doi: 10.48550/arXiv.1911.05248. URL https://arxiv.org/abs/1911.05248.

Yen-Chang Hsu, Ting Hua, Sungen Chang, Qian Lou, Yilin Shen, and Hongxia Jin. Language model compression with weighted low-rank factorization. In *The Tenth International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2207.00112.

Toni J. B. Liu, Jiajun Bao, Yizhou Liu, Gurbir Arora, Nicolas Boullé, Raphaël Sarfati, and Christopher J. Earls. The geometry of ignorance: LLMs know when to temper Bayesian priors. *arXiv preprint arXiv:2609.02959*, 2026. doi: 10.48550/arXiv.2609.02959. URL https://arxiv.org/abs/2609.02959.

Iuri Macocco, Nora Graichen, Gemma Boleda, and Marco Baroni. Not a nuisance but a useful heuristic: Outlier dimensions favor frequent tokens in language models. In *Proceedings of the 8th BlackboxNLP Workshop*, pp. 109–136, 2025. URL https://aclanthology.org/2025.blackboxnlp-1.6/.

Clara Meister, Wojciech Stokowiec, Tiago Pimentel, Lei Yu, Laura Rimell, and Adhiguna Kuncoro. A natural bias for language generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 243–255. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-short.22. URL https://aclanthology.org/2023.acl-short.22/.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Kiho Park, Todd Nief, Yo Joong Choe, and Victor Veitch. The information geometry of softmax: Probing and steering. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. doi: 10.48550/arXiv.2602.15293. URL https://arxiv.org/abs/2602.15293.

Charbel Sakr and Brucek Khailany. ESPACE: Dimensionality reduction of activations for model compression. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0556. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/1f6591cc41be737e9ba4cc487ac8082d-Abstract-Conference.html.

Samuel Salfati. Variance is not importance: Structural analysis of transformer compressibility across model scales. *arXiv preprint arXiv:2604.20682*, 2026. doi: 10.48550/arXiv.2604.20682. URL https://arxiv.org/abs/2604.20682.

Nilesh Sarkar and Dawar Jyoti Deka. Causal dimensionality of transformer representations: Measurement, scaling, and layer structure. *arXiv preprint arXiv:2605.08740*, 2026. doi: 10.48550/arXiv.2605.08740. URL https://arxiv.org/abs/2605.08740.

Rickmer Schulte and David Rügamer. Rethinking intrinsic dimension estimation in neural representations. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 10–18. PMLR, 2026. URL https://proceedings.mlr.press/v300/schulte26a.html.

Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. Beyond variance: Knowledge-aware LLM compression via Fisher-aligned subspace diagnostics. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 47938–47954. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.2214. URL https://aclanthology.org/2026.acl-long.2214/.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for BERT model compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 4323–4332. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1441. URL https://aclanthology.org/D19-1441/.

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4527–4546, 2021. URL https://aclanthology.org/2021.emnlp-main.372/.

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, 2023. doi: 10.52202/075280-2230. URL https://papers.nips.cc/paper_files/paper/2023/hash/a0e66093d7168b40246af1cddc025daa-Abstract-Conference.html.

Sihan Wang, Jiayi Zhao, Qingyan Cao, Hongbo Yao, and Lin Shu. FishBack: Pullback Fisher geometry for optimal activation steering in transformers. *arXiv preprint arXiv:2605.17231*, 2026. doi: 10.48550/arXiv.2605.17231. URL https://arxiv.org/abs/2605.17231.

A Linear Reduction Methods

The four methods in Section 3.1 use the same mean-preserving reconstruction. They differ in how they choose the retained directions. This appendix gives their constructions; Appendix C specifies the fitting data, optimisation settings and validation procedure.

A.1 Mean-Preserving Reconstruction and PCA

Let $h(c_i) \in \mathbb{R}^D$ be the post-normalization output state for fitting prefix c_i, for $i = 1, \ldots, N$. We centre these states using their mean:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} h(c_i), \qquad x_i = h(c_i) - \mu.$$

At rank k, every method reconstructs the state as

$$R_k(h) = \mu + A_k(h - \mu), \qquad \operatorname{rank}(A_k) \leq k.$$

The mean is preserved exactly. The rank therefore limits variation around the mean.

Ordinary PCA chooses

$$A_k = V_k V_k^\top,$$

where the columns of V_k are the k leading principal directions of the centred fitting states. This map minimises squared reconstruction error,

$$\sum_{i=1}^{N} \|x_i - A_k x_i\|_2^2,$$

among maps of rank at most k. It uses activation variance, without considering how reconstruction errors affect the output distribution.

A.2 Output-Fisher-Weighted PCA

Output-Fisher-weighted PCA accounts for the effect of state errors on predictions. It uses the softmax Fisher geometry (Park et al., 2026) in an activation-space construction related to Fisher-weighted low-rank approximation (Hsu et al., 2022).

For prefix c, write

$$p = p_\theta(\cdot \mid c) = \mathrm{softmax}(W_U h(c) + b)\,,$$

where W_U is the unembedding matrix and $b = 0$ for models without an output bias. The output Fisher matrix is

$$G(c) = W_U^\top \left[\mathrm{diag}(p) - pp^\top\right] W_U.$$

For a small state perturbation δ,

$$D_{\mathrm{KL}}(p_\theta(\cdot \mid c) \,\|\, \mathrm{softmax}(W_U[h(c)+\delta]+b)) = \frac{1}{2}\delta^\top G(c)\delta + O(\|\delta\|^3).$$

Thus, the Fisher matrix weights state errors according to their local effect on the prediction. An output bias changes this metric through p.

For this method, we replace the prefix-specific matrices by one average over $N_G = 4096$ sampled fitting prefixes:

$$\overline{G}_{\mathrm{fit}} = \frac{1}{N_G}\sum_{i=1}^{N_G} G(c_i).$$

We symmetrise this matrix and compute

$$\overline{G}_{\mathrm{fit}} = Q\Lambda Q^\top.$$

To stabilise the inverse, we apply an eigenvalue floor:

$$\lambda_r^\epsilon = \max(\lambda_r, \epsilon), \qquad \epsilon = 10^{-8}.$$

Writing $\Lambda_\epsilon = \mathrm{diag}(\lambda_1^\epsilon, \ldots, \lambda_D^\epsilon)$, we define

$$\overline{G}_{\mathrm{fit},\epsilon} = Q\Lambda_\epsilon Q^\top, \qquad S = \overline{G}_{\mathrm{fit},\epsilon}^{1/2}, \qquad S^{-1} = \overline{G}_{\mathrm{fit},\epsilon}^{-1/2}.$$

We then transform the centred states,

$$y_i = Sx_i,$$

and apply ordinary PCA to the transformed states. If U_k contains their k leading principal directions, the reconstruction map is

$$A_k = S^{-1}U_kU_k^\top S.$$

The multiplication by S weights the states, $U_kU_k^\top$ retains k directions, and S^{-1} returns the result to the original coordinates.

This map solves

$$A_k \in \operatorname*{arg\,min}_{\mathrm{rank}(A)\le k} \sum_{i=1}^{N} \big((I-A)x_i\big)^\top \overline{G}_{\mathrm{fit},\epsilon}\big((I-A)x_i\big).$$

Indeed, the change of coordinates $y_i = Sx_i$ turns this objective into ordinary squared reconstruction error.

This construction uses both activation variation and output sensitivity. It is not equivalent to retaining the leading eigenvectors of the Fisher matrix alone. Its optimality also concerns the fixed quadratic objective above, not KL itself.

A.3 Paired-Fisher and KL Training

The two trained methods learn a rank-k map directly:

$$A_k = UV^\top, \qquad U, V \in \mathbb{R}^{D\times k}.$$

The factors are initialised from either PCA or output-Fisher-weighted PCA. The mean μ remains fixed.

For fitting prefix c_i, the reconstruction removes the state component

$$e_k(c_i) = h(c_i) - R_k(h(c_i)).$$

Paired-Fisher training minimises

$$\frac{1}{2N}\sum_{i=1}^{N} e_k(c_i)^\top G(c_i) e_k(c_i).$$

Unlike output-Fisher-weighted PCA, it evaluates each reconstruction error using the Fisher matrix of its own prefix, rather than one corpus-average matrix.

KL training instead minimises the actual change in the output distributions:

$$\frac{1}{N}\sum_{i=1}^{N} \Delta_k(c_i), \qquad \Delta_k(c_i) = D_{\mathrm{KL}}(p_\theta(\cdot \mid c_i) \,\|\, p_k(\cdot \mid c_i)) \,.$$

All four methods are fitted without changing the model or its output head. Their final comparison uses KL on held-out prefixes, regardless of the objective used for fitting.

B Proof of Proposition 1

Proposition 1 shows that nearly one-dimensional sensitivity at each prefix can coexist with a corpus-average Fisher matrix whose participation rank approaches the full model width.

For prefix c_i, let $p_i = p_\theta(\cdot \mid c_i)$. Its output Fisher matrix is

$$G_i = W_U^\top \left[\mathrm{diag}(p_i) - p_i p_i^\top\right] W_U.$$

For a nonzero positive semidefinite matrix M, its participation rank is

$$\mathrm{PR}(M) = \frac{\mathrm{tr}(M)^2}{\mathrm{tr}(M^2)}.$$

Proposition 1 (restated). For every width $D \geq 1$ and every $\eta > 0$, there is a biasless, full-support softmax model with D prefixes such that every local Fisher matrix has exact rank D, while

$$\mathrm{PR}(G_i) < 1 + \eta \quad \text{for every } i, \qquad \mathrm{PR}\left(\frac{1}{D}\sum_{i=1}^{D} G_i\right) > D - \eta.$$

Proof. We construct a model whose sensitivity concentrates along a different coordinate for each prefix. Averaging over the prefixes then distributes sensitivity across all coordinates.

Use $D+1$ tokens, indexed by $0, \ldots, D$, and no output bias. Let the rows of the unembedding matrix be

$$w_0 = 0, \qquad w_j = e_j \quad (j = 1, \ldots, D),$$

where e_j is the j-th standard basis vector of $\mathbb{R}^D$.

First, every full-support output distribution gives a Fisher matrix of exact rank D. To see this, let

$$G_p = W_U^\top [\mathrm{diag}(p) - pp^\top] W_U.$$

For any $v \in \mathbb{R}^D$,

$$v^\top G_p v = \mathrm{Var}_{Y\sim p}((W_U v)_Y) \,.$$

Because every token has positive probability, this variance is zero only when all entries of $W_U v$ are equal. But

$$W_U v = (0, v_1, \ldots, v_D),$$

so this happens only when $v = 0$. Hence G_p is positive definite and has rank D.

Now choose a positive scalar M. For prefix c_i, set

$$h(c_i)_i = 0, \qquad h(c_i)_j = -M \quad (j \neq i).$$

Table 2: Models in the main experiments.

Family	Models	Widths D
GPT-2	GPT-2; GPT-2 Medium	768; 1,024
Pythia	70M; 160M; 410M; 1.4B; 2.8B; 6.9B	512; 768; 1,024; 2,048; 2,560; 4,096
Qwen2.5	0.5B; 1.5B; 3B; 7B	896; 1,536; 2,048; 3,584
Qwen3	0.6B	1,024
Qwen3.5	0.8B; 2B; 4B	1,024; 2,048; 2,560
SmolLM2	360M	960
SmolLM3	3B	2,048
LFM2	350M	1,024
Gemma 4	E2B; E4B	1,536; 2,560

At every finite M, all softmax probabilities are positive, so every G_i retains exact rank D.

As $M \to \infty$, the prediction at prefix c_i assigns probability $1/2$ to token 0, probability $1/2$ to token i, and vanishing probability to the other tokens. Therefore,

$$G_i \longrightarrow \frac{1}{4} e_i e_i^\top .$$

Each local Fisher matrix consequently has participation rank approaching 1. Their average satisfies

$$\frac{1}{D} \sum_{i=1}^{D} G_i \longrightarrow \frac{1}{4D} I_D,$$

whose participation rank is D.

Participation rank is continuous at these nonzero limits. Since there are finitely many prefixes, a sufficiently large finite M gives both required inequalities while preserving exact rank D for every local matrix. □

The proposition separates local and corpus-average participation ranks. It does not establish a lower bound on the reconstruction rank needed at a nonzero KL tolerance.

C Experimental Protocol

All scripts needed to reproduce the experiments will be released in a public repository, with the link withheld to preserve anonymity during review.

C.1 Models

The main experiments use 21 models from nine families, with model widths from 512 to 4,096. The exact model version is recorded for every experiment.

Table 2 gives the models and widths used in the main experiments.

C.2 Data and Splits

The main corpus is the training split of WikiText-103 (`Salesforce/wikitext`, `wikitext-103-v1`). Additional datasets are FineWeb-Edu (`HuggingFaceFW/fineweb-edu`), the full Pile validation set, and its GitHub documents (`mit-han-lab/pile-val-backup`). For each corpus, we retain the first 450 documents containing at least 2,500 characters and shuffle them with a fixed random seed. We assign 300 documents to fitting, 75 to validation, and 75 to testing. We take at most three non-overlapping blocks of 513 tokens from each document. No block crosses a document boundary. We define token-frequency groups from the validation split.

C.3 MAP FITTING

The mean state and PCA basis are computed from all tokens in the fitting set. We estimate the output Fisher metric from 4,096 states in that set, sampled with seed 0.

We train each objective twice, starting once from PCA and once from output-Fisher-weighted PCA. We optimize each map for at most 800 Adam steps, sampling 64 fitting states with replacement at each step. The learning rate is 10^{-4}, and gradients are clipped. Every 25 steps, we evaluate the training objective on 4,096 fixed validation states. Training stops after 12 consecutive checks without an improvement greater than 10^{-6}. We retain the step with the lowest validation value. The validation objective is KL for KL-trained maps and the paired-Fisher objective for paired-Fisher maps.

We test ranks obtained from the following fractions of model width:

$$\left\{\tfrac{1}{2}, \tfrac{2}{3}, 0.70, 0.75, 0.80, \tfrac{5}{6}, \tfrac{7}{8}, \tfrac{11}{12}, \tfrac{23}{24}, 1\right\} D.$$

C.4 EVALUATION

After selecting the map on the validation set, we apply it to every token in the 75 test documents. The reported KL is the average across these tokens. To estimate uncertainty, we resample the 75 documents with replacement 1,000 times and recompute the average. The middle 95% of these values gives the reported interval.

We also test whether the result depends on which articles are selected. For GPT-2, Pythia-410M, and Qwen2.5-0.5B, we repeat the complete fitting, validation, and testing procedure on three non-overlapping sets of 450 WikiText articles. Resampling token blocks instead of documents gives the same conclusions.

Models run in float32 except for Pythia-6.9B, Qwen2.5-7B, and Gemma-4-E4B, which run in bfloat16. We fit all maps in float32. Runs used an NVIDIA H100 or an Apple M4 Max.

D ADDITIONAL WIDTH REQUIREMENT RESULTS

D.1 REQUIREMENT ACROSS DEPTH

Table 3 gives the individual model results for the depth experiment in Section 3.3. Intermediate layers often tolerate lower ranks than the output, but the pattern varies between models. It shows exceptions to the general trend: Pythia-6.9B requires more width at its middle layers than at its output. States whose norms exceed five times the median within their block are left unchanged. The experiment uses one document split and one random seed.

D.2 SHARED-RANK CONTROLS

We ask whether the high measured requirement comes from the fitting method, the selected articles, or the constraint that the reconstruction preserve the mean.

We compare the four fitting methods on GPT-2, Pythia-410M, and Qwen2.5-0.5B. We repeat fitting and evaluation on three non-overlapping sets of 450 WikiText articles. Separately, across all 21 models, we allow the reconstruction to learn its constant offset instead of preserving the fitting mean.

Table 4 reports the method comparison and the variation across article sets. At half width, KL training gives the lowest distortion: 0.112, 0.208, and 0.229 nats for the three models, respectively. Changing the articles changes KL by 0.001–0.004 nats at each rank. Learning the offset changes KL by at most 0.003 nats across the 21 models.

Direct KL training improves the reconstruction, but all four methods remain above the 0.0488 tolerance at half width. Across article sets, the first passing rank stays the same or moves by one tested step. Allowing a learned offset has little effect.

These choices do not explain away the high measured requirement. It persists across the tested methods, article samples, and offset choices.

Table 3: Requirement across depth. The first block gives the first passing fraction for PCA and KL training; the second gives PCA KL at two-thirds width.

	First passing fraction: PCA / KL				KL at two-thirds width: PCA			
Model	1/4	1/2	3/4	exit	1/4	1/2	3/4	exit
GPT-2	.67 / .42	.67 / .67	.70 / .67	.80 / .70	.018	.046	.049	.090
GPT-2 Medium	.67 / .50	.67 / .67	.67 / .67	.80 / .70	.023	.042	.048	.091
Pythia-70M	.80 / .75	.83 / .80	.83 / .80	.83 / .75	.114	.180	.134	.127
Pythia-160M	.75 / .67	.80 / .70	.80 / .70	.80 / .75	.069	.088	.090	.080
Pythia-410M	.67 / .67	.67 / .67	.75 / .70	.88 / .80	.030	.042	.066	.148
Pythia-1.4B	.50 / .50	.67 / .67	.75 / .75	.83 / .75	.020	.031	.073	.112
Pythia-2.8B	.50 / .42	.67 / .67	.80 / .80	.80 / .75	.017	.029	.099	.100
Pythia-6.9B	.75 / .75	.83 / .83	.83 / .83	.75 / .70	.053	.075	.104	.072
Qwen2.5-0.5B	.67 / .50	.50 / .42	.67 / .67	.88 / .83	.025	.015	.048	.167
Qwen2.5-1.5B	.67 / .50	.50 / .50	.67 / .50	.88 / .80	.024	.019	.024	.161
Qwen2.5-3B	.67 / .67	.50 / .42	.50 / .50	.88 / .80	.038	.016	.019	.154
Qwen2.5-7B	.50 / .50	.67 / .67	.50 / .50	.80 / .75	.021	.025	.018	.093
Qwen3-0.6B	.67 / .67	.50 / .50	.67 / .67	.83 / .80	.029	.018	.034	.129
SmolLM2-360M	.67 / .67	.50 / .42	.67 / .67	.88 / .83	.030	.017	.028	.175
LFM2-350M	.67 / .67	.67 / .67	.75 / .75	.83 / .80	.047	.028	.074	.112
Qwen3.5-0.8B	.50 / .42	.42 / .42	.70 / .67	.80 / .75	.014	.011	.050	.105
Qwen3.5-2B	.42 / .33	.42 / .33	.67 / .67	.75 / .67	.010	.009	.037	.069
Qwen3.5-4B	.42 / .33	.33 / .25	.67 / .67	.70 / .67	.011	.007	.035	.053
Gemma-4-E2B	.67 / .67	.50 / .42	.67 / .67	.80 / .80	.035	.026	.030	.121
Gemma-4-E4B	.67 / .67	.92 / .92	.67 / .67	.80 / .80	.032	.327	.027	.098
SmolLM3-3B	.42 / .33	.42 / .33	.75 / .75	.88 / .80	.012	.010	.065	.144

Table 4: Method and split controls. The first four columns give mean KL at half width. The last two give the first passing ranks across three article pools; "Interval" uses the bootstrap upper bound.

Model	PCA	Fisher PCA	Paired Fisher	KL training	Point crossing	Interval crossing
GPT-2	.227	.169	.116	.112	.67–.70	.70
Pythia-410M	.322	.278	.217	.208	.80	.80–.83
Qwen2.5-0.5B	.358	.310	.239	.229	.83	.83

D.3 Training Stage

We test whether the high shared requirement develops during pretraining and whether it changes alongside local Fisher rank. If falling local rank indicated that states were becoming easier to reconstruct, the shared requirement should fall too.

Figure 5 reports measurements at 20 checkpoints from three Pythia models. At each checkpoint, we measure prediction loss, Fisher rank for individual prefixes and for the corpus, and shared requirement. Each checkpoint is measured once. The shared requirement rises while the Fisher rank around individual predictions falls.

In later checkpoints of SmolLM3-3B, the shared requirement remains at 80% of model width between 3.8×10^{11} and 1.1×10^{13} training tokens, while the corpus Fisher rank falls from 47% to 30%. These checkpoints cover later training and do not establish whether SmolLM3 follows the early trajectory observed in Pythia.

A falling Fisher rank during training does not necessarily mean that a narrower shared reconstruction can preserve predictions.

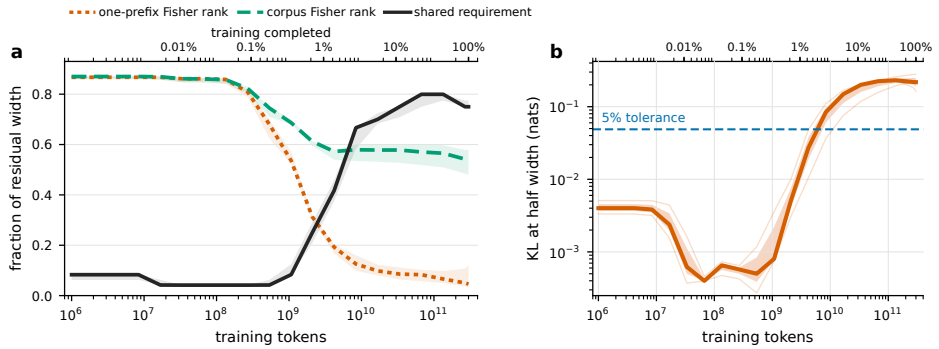

Figure 5: Local and shared dimensions during Pythia pretraining. **(a)** One-prefix Fisher rank, corpus Fisher rank, and shared requirement. **(b)** KL at half width. Lines and bands show cross-model medians and interquartile ranges; faint lines show individual models. Each checkpoint is measured once.

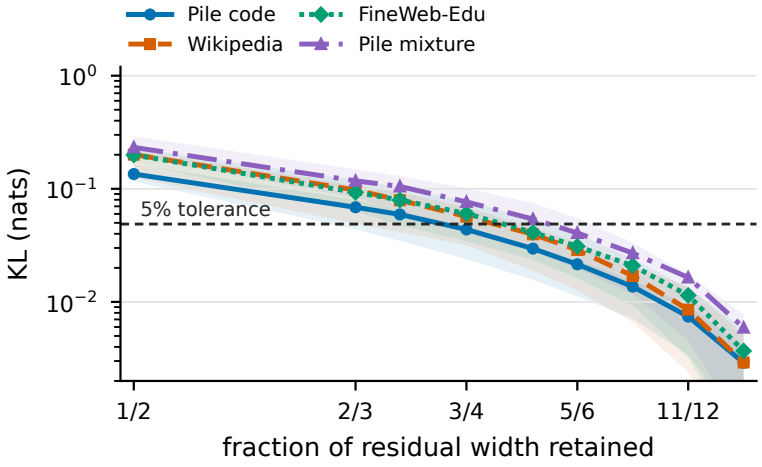

Figure 6: KL by retained width on four text distributions. Curves show medians over the nine models; bands show interquartile ranges. Each map is fitted and tested on disjoint documents from the same corpus.

D.4 Text Distribution

We test whether the high shared requirement is specific to WikiText-103 by repeating the measurement on other text distributions.

Figure 6 compares the KL curves of nine models on Pile code, WikiText-103, FineWeb-Edu and the mixed Pile. For each model and corpus, we fit the reconstruction on one set of documents and evaluate it on separate documents from the same corpus. Pile code has the lowest or tied lowest requirement, while the mixed Pile has the highest.

Similar required ranks would not establish that the same reconstruction works across corpora. We therefore test whether a map fitted on one corpus preserves predictions on another without refitting.

Table 5 reports transfer between WikiText-103, Pile code and the mixed Pile. We apply each fitted map to the other corpora without refitting it. All 11 transfers that meet the tolerance at five-sixths width use maps fitted on the mixed Pile.

The required rank depends on the text distribution, and a reconstruction fitted on one corpus does not reliably preserve predictions on another.

Table 5: Cross-corpus transfer. Each cell gives KL at $5D/6$ and, in parentheses, the first passing rank. Rows are fitting corpora and columns are test corpora. W, M, and C denote Wikipedia, the mixed Pile, and Pile code.

Fit	W	M	C
GPT-2			
W	.015 (.70)	.053 (.88)	.069 (.88)
M	.028 (.80)	.020 (.75)	.013 (.67)
C	.118 (.92)	.062 (.88)	.007 (.67)
Pythia-410M			
W	.034 (.80)	.116 (.92)	.186 (.96)
M	.079 (.92)	.058 (.88)	.045 (.83)
C	.325 (1.00)	.194 (.96)	.022 (.75)
Qwen2.5-0.5B			
W	.038 (.83)	.174 (.96)	.355 (1.00)
M	.093 (.92)	.066 (.88)	.061 (.88)
C	.327 (1.00)	.218 (.96)	.033 (.80)
Qwen3.5-0.8B			
W	.013 (.75)	.091 (.92)	.226 (.96)
M	.028 (.80)	.024 (.80)	.023 (.80)
C	.150 (.92)	.110 (.92)	.014 (.70)
Qwen3.5-2B			
W	.008 (.67)	.080 (.92)	.218 (.96)
M	.019 (.80)	.016 (.75)	.018 (.75)
C	.117 (.92)	.094 (.92)	.011 (.67)

Fit	W	M	C
Qwen3.5-4B			
W	.005 (.67)	.059 (.88)	.153 (.92)
M	.012 (.75)	.010 (.70)	.012 (.75)
C	.084 (.88)	.068 (.88)	.009 (.67)
Gemma-4-E2B			
W	.036 (.80)	.083 (.92)	.104 (.96)
M	.120 (.92)	.047 (.83)	.035 (.83)
C	.169 (1.00)	.110 (.96)	.023 (.75)
Gemma-4-E4B			
W	.031 (.80)	.118 (.96)	.185 (1.00)
M	.098 (.92)	.041 (.83)	.031 (.80)
C	.332 (1.00)	.227 (1.00)	.022 (.75)
SmolLM3-3B			
W	.029 (.80)	.181 (.96)	.376 (1.00)
M	.083 (.88)	.055 (.88)	.049 (.88)
C	.361 (1.00)	.238 (1.00)	.025 (.80)

D.5 Target-Frequency Groups

We test whether the high corpus-wide requirement comes only from prefixes followed by infrequent tokens, or whether it also holds for prefixes followed by frequent tokens.

We group test prefixes according to whether their observed next token belongs to the 100 most frequent types or falls outside the top 1,000. We evaluate the same fitted reconstructions separately on each group. KL compares the full next-token distributions at these positions, rather than only the probability of the observed token. Table 6 reports the smallest passing rank for each group, with results grouped by model family.

Table 6: First passing ranks by target frequency, reported as ranges within each model family.

Family	Models	Top 100	Outside 1,000
GPT-2	2	.67	.80
Pythia	6	.50–.67	.80–.88
Qwen2.5	4	.67–.70	.83–.88
Qwen3	1	.67	.88
SmolLM2	1	.70	.92
LFM2	1	.67	.83
Qwen3.5	3	.50–.67	.70–.80
Gemma 4	2	.67	.88
SmolLM3	1	.70	.88

Prefixes followed by infrequent tokens require higher ranks, but even those followed by frequent tokens usually need about two-thirds of model width. The high shared requirement is therefore not confined to infrequent-token positions.

E CONTEXT-SENSITIVITY CONTROLS

E.1 PREFIX-SPECIFIC RECONSTRUCTION ERRORS

The experiment in Section 4.2 uses 103,424 test positions for GPT-2, 103,424 for Pythia-410M and 103,936 for Qwen2.5-0.5B. At half width, we estimate the effect of each removed state vector using its prefix's Fisher matrix. We then keep the vector fixed and replace the matrix with one from another prefix. We test both replacements with similar entropy and target-token frequency, and unrestricted random replacements. Each replacement control is repeated three times.

For GPT-2, Pythia-410M and Qwen2.5-0.5B, respectively, the original matrices give mean estimates of 0.168, 0.287 and 0.329 nats. Matrices from prefixes with similar entropy and token frequency give 0.050, 0.089 and 0.101 nats. The Spearman correlation with measured KL is 0.94–0.998 for the original matrices, 0.64–0.79 for similar-prefix replacements and 0.28–0.40 for unrestricted replacements.

The same removed vector therefore has a larger estimated effect under its own prefix's Fisher matrix, even when the replacement prefix has similar entropy and token frequency.

E.2 LOCAL-SUBSPACE OVERLAP AND TARGET DIVERSITY

Figure 7 compares the eight leading Fisher directions between prefixes across 21 models. An overlap score of 1 means identical subspaces. The mean score is 0.34 for prefixes followed by the same token and 0.12 for different tokens in the same frequency group. Both exceed the random-subspace baseline. Token identity therefore influences the sensitive directions, but does not determine them.

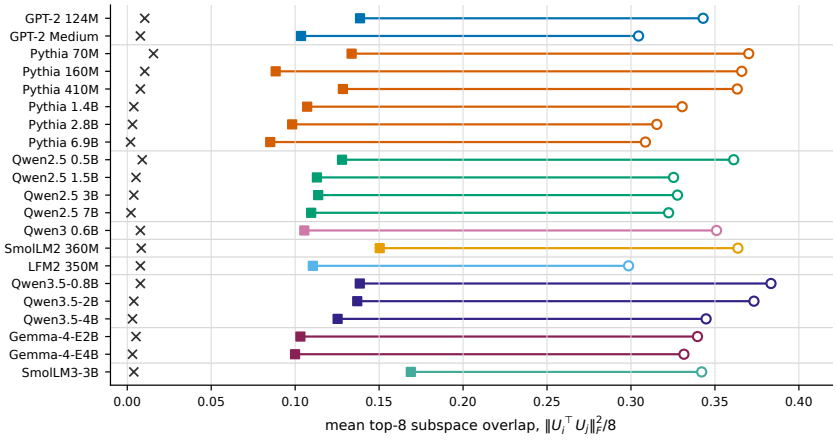

Figure 7: Top-eight local Fisher subspace overlap across 21 models. Hollow circles compare contexts with the same target token; squares compare contexts with different target tokens in the same frequency class; crosses show random-subspace overlap. Lines join measurements from the same model. Mean same-target and different-target overlap is 0.34 and 0.12, respectively.

Figure 8 holds the number of prefixes fixed within each model, between 64 and 256, while varying the number of observed next-token types. Combining more token types generally increases the number of directions needed to capture 90% of the average Fisher sensitivity. This growth therefore persists without increasing the number of prefixes being averaged.

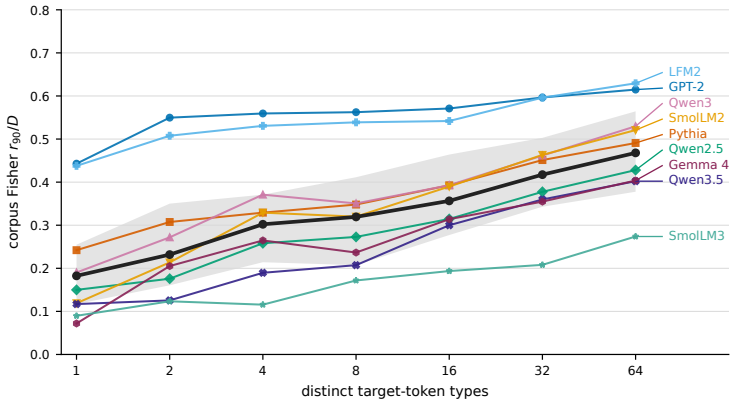

Figure 8: Corpus-average output Fisher rank as target-token diversity increases. Each coloured line is a family median; the black line and grey band show the median and interquartile range across all 21 models. The context count is fixed within each model (64–256 depending on width). Values are fractions of model width.

F Generation Protocol and Results

F.1 Matched Controls

We test whether changes in uncertainty or output probabilities can reproduce the frequency shift caused by low-rank reconstruction.

For each of the 21 models, we generate 128 tokens from the same 200 held-out WikiText-103 prefixes of 64 tokens. We use nucleus sampling ($p = 0.9$, temperature 1) and three random seeds shared across interventions. The retained fractions are $1/16$, $1/8$, $1/4$, $1/2$, and $5/6$. Token frequencies are estimated from the WikiText-103 validation stream.

For each model and rank, we calibrate two controls on held-out text before generation. The temperature control adjusts the intact model's temperature until its average output entropy matches that of the low-rank reconstruction. The noise control adds random noise to the final state, with equal variance in every direction. We choose its scale to match the low-rank reconstruction's average KL from the intact model.

These settings remain fixed during generation. Entropy and KL are matched on the calibration text, not on the different prefixes subsequently generated by each intervention.

Table 7 reports the median frequency changes at half width across the 21 models.

Table 7: Changes at half width, in percentage points relative to intact generation.

Intervention	Outside top 1,000	Inside top 100
Low-rank reconstruction	-9.3	$+9.6$
Entropy-matched temperature	$+0.4$	-0.9
KL-matched isotropic noise	$+3.0$	-3.7

Low-rank reconstruction shifts generation towards frequent tokens. Temperature has little effect, while noise shifts generation towards infrequent tokens. Matching average entropy or KL on held-out text therefore does not reproduce the frequency shift.

F.2 Repetition

We also check whether the frequency shift accompanies increasingly repetitive text. For GPT-2, Pythia-410M and Qwen2.5-0.5B, we measure the fraction of continuations containing a sequence

of four tokens at least three times. We also report Distinct-4: the number of distinct four-token sequences divided by their total count, pooled across the generated continuations.

Table 8 compares intact generation with reconstruction at half width.

Table 8: Repetition and Distinct-4, shown as intact $\rightarrow$ half width.

Model	Repetitive continuations	Distinct-4
GPT-2	$0.35 \rightarrow 0.28$	$0.86 \rightarrow 0.90$
Pythia-410M	$0.41 \rightarrow 0.42$	$0.86 \rightarrow 0.84$
Qwen2.5-0.5B	$0.57 \rightarrow 0.56$	$0.78 \rightarrow 0.79$

At half width, repetition decreases in GPT-2 and changes little in Pythia and Qwen. The frequency shift therefore occurs without a consistent increase in repetition in these three models.

At $D/16$, however, Pythia and Qwen become strongly repetitive. Their frequency shifts at this rank cannot be interpreted separately from this deterioration in generation.