Gradient Starvation in Low-Rank Language Model Pretraining

Anonymous ACL submission

Abstract

Low-rank pretraining consistently underperforms full-rank (FR) pretraining, but its representational and optimisation mechanisms remain unclear. We present a mechanistic analysis using GPT-2 on BabyLM dataset across multiple ranks. Two findings stand out. First, the FR–LR representational split emerges before substantial learning begins: the rank constraint separates models in representation space, and gradient starvation then sustains and amplifies the split. We observed that the FR effective gradient dimensionality expands during training while LR effective dimensionality remains bounded. Second, optimisation success and representational alignment are dissociated; this was materialised in that spectral normalisation improves perplexity substantially without improving FR–LR similarity. Closing the performance gap does not imply closing the representational gap. Our findings suggest efficient pretraining methods should be evaluated not only by downstream performance, but also by the representational and optimisation trajectories through which performance emerges.

1 Introduction

Efficient language-model pretraining increasingly relies on low-dimensional training constraints. Low-rank parametrisations reduce memory and compute by restricting weight matrices to a factorised form $W = BA$ where $r \ll \min(m, n)$ (Hu et al., 2022; Lialin et al., 2024). Gradient projection methods compress optimiser states while preserving full-rank weights (Zhao et al., 2024), while structured accumulation schemes recover expressivity through periodic restarts (Lialin et al., 2024; Li et al., 2025). Together these approaches have made low-rank pretraining a practical choice for resource-constrained settings (Warstadt et al., 2023; Hu et al., 2024; Wei et al., 2024).

Yet the performance gap with full-rank (FR) pretraining persists across methods and scales (Lialin et al., 2024; Zhao et al., 2024; Diehl Martinez et al., 2024). The field's response has been almost entirely engineering: design a better method that narrows the gap. What has received far less attention is the mechanistic question: *what does the rank constraint do to the representations a language model learns, and why does the damage resist correction?*

This paper provides evidence for a mechanistic account of low-rank (LR) pretraining dynamics in language models. Using GPT-2 pretrained on the BabyLM Strict-Small corpus (Hu et al., 2024) as our primary setting, we conduct a systematic analysis organised around six objectives spanning divergence timing, stabilisation mechanism, gradient dimensionality, spectral causality, performance scaling, and recoverability. We use Centered Kernel Alignment (CKA; Kornblith et al., 2019) to track representational divergence, effective rank (Roy and Vetterli, 2007) to characterise gradient dimensionality, and targeted spectral interventions to establish causality. We complement language findings with a parallel ViT (Dosovitskiy et al., 2021) analysis on CIFAR-10 as a cross-domain check, not an equal headline result.

The split is structural, not learned. At 8,192 tokens, approximately one batch, LR models at all ranks are already separated from FR in representation space. The rank constraint alone, before meaningful learning has occurred, places models in different representational regimes. This shifts the central question from *when* training diverges to *why* initialisation separates the two.

Gradient starvation is the sustaining mechanism. We identify a mechanism that explains much of this gap: gradient effective rank starvation. In language, FR's gradient effective rank grows from 306 to 510 over training as the model discovers new optimisation directions, while LR's remains bounded by the rank constraint. Qualitatively, unlike vision, where the ratio is approximately constant, the gap *widens*.

The rank constraint imposes a ceiling on gradient dimensionality that FR actively surpasses, and the longer training runs, the further LR falls behind.

Stabilisation reveals a language-specific pathology. All LR language runs stabilise by *exhaustion*: representations drift far from initialisation before step-to-step changes become small, indicating the model has run out of productive gradient directions. FR language models also stabilise by exhaustion which is unlike FR vision models, whose shallow layers show clean convergence. Two independently trained FR language models *diverge* in representation space over training rather than converging. The rank constraint amplifies a pathology already latent in language pretraining.

Spectral geometry is causal at initialisation, not throughout. Intervening at the first checkpoint t^* to flatten the LR model's singular value spectrum raises CKA alignment by 288% and reduces perplexity from 490 to 188 (61% improvement), confirming that spectral geometry at initialisation causally drives the early split. However, applying this correction continuously throughout training does not sustain the alignment gains: gradient flow overwrites the spectral initialisation within a few updates, and task performance improves without improving representational alignment with FR. This dissociation identifies gradient dimensionality as the dominant ongoing constraint.

The gap is only partially recoverable. Full fine-tuning of a trained LR model closes only 12.5–42% of the performance gap in language depending on training budget; reinitialising all transformer blocks before fine-tuning recovers up to 73%, yet a substantial residual persists. FR and LR pretraining lead to different basins in parameter space (Lyle et al., 2023; Dohare et al., 2024), and post-hoc adaptation cannot fully escape them.

These findings have implications beyond the efficiency literature. Work on transfer learning (Aghajanyan et al., 2021), representation probing (Diehl Martinez et al., 2024), and model analysis (Kornblith et al., 2019) assumes pretraining produces a representational substrate of known quality. If downstream representations matter, as they do across NLP (Devlin et al., 2019; Brown et al., 2020), the rank of pretraining should be treated as a first-class experimental variable.

Contributions.

1. The FR–LR representational split in language models is present before any meaningful learning, implicating the rank constraint itself rather than learning dynamics as the primary source of divergence.

2. Gradient effective rank starvation sustains the gap: a widening deficit in gradient dimensionality the optimiser cannot compensate for, qualitatively more severe and dynamic in language than in vision.

3. Spectral geometry at initialisation causally drives the initial split, but continuous spectral correction cannot overcome the gradient dimensionality constraint during training (Roy and Vetterli, 2007; Kornblith et al., 2019).

4. The FR–LR gap is only partially recoverable by fine-tuning, confirming the two pretraining regimes occupy different basins in parameter space (Lyle et al., 2023; Dohare et al., 2024).

5. Cross-domain validation on ViT/CIFAR-10 (Dosovitskiy et al., 2021) shows core phenomena replicate while revealing systematic domain differences that illuminate the unique challenges of language pretraining.

6. All code, training scripts, probing pipelines, and collated results are released to support reproducibility.[1]

2 Background

Low-rank pretraining constrains each weight matrix $W \in \mathbb{R}^{m\times n}$ to a factorised form $W = BA$, where $B \in \mathbb{R}^{m\times r}$, $A \in \mathbb{R}^{r\times n}$, and $r \ll \min(m, n)$, reducing trainable parameters from mn to $r(m + n)$ (Lialin et al., 2024; Zhao et al., 2024). The performance cost of this constraint is well documented: LR pretraining consistently underperforms FR pretraining at low rank (Lialin et al., 2024; Zhao et al., 2024; Wei et al., 2024).

Prior work offers partial clues about why this occurs. Diehl Martinez et al. (2024) show that layers with lower effective rank in small language models exhibit slower and less stable convergence during pretraining. Godey et al. (2024) establish a theoretical lower bound on achievable loss when model width is insufficient for the rank of the target output distribution. However, neither work characterises *when* the representational divergence between FR

[1] [anonymised for review].

and LR arises, what sustains it during training, or whether it is reversible. Intuitively, if FR training can explore many independent update directions while LR training cannot, the two models may diverge functionally even under identical optimisation settings. This paper provides evidence for a mechanistic account of that divergence.

Our focus differs from prior work. Where previous work asks how to close the performance gap, we ask what creates and sustains it at the representational and optimisation level. This distinction motivates the mechanistic tools introduced below.

2.1 Metrics

Centered Kernel Alignment (CKA). We measure representational divergence using linear CKA (Kornblith et al., 2019), which is invariant to orthogonal transformations and isotropic scaling and therefore well-suited to comparing representations across independently trained models with no a priori alignment. For activation matrices $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^{n\times q}$ collected over n examples:

$$\mathrm{CKA}(X,Y) = \frac{\|Y^\top X\|_F^2}{\|X^\top X\|_F \, \|Y^\top Y\|_F}. \quad (1)$$

We report three variants: *cross-run CKA* between an LR model and a fixed FR reference (FR–LR divergence); *self-CKA* between consecutive checkpoints of the same run (representational velocity); and *init-CKA* between a checkpoint and its own initialisation (cumulative drift). All CKA values are means across transformer layers unless stated otherwise; layer-wise analyses showing consistent trends are provided in Appendix G.

Gradient effective rank. We characterise gradient dimensionality using effective rank (Roy and Vetterli, 2007), the exponential Shannon entropy of the normalised singular value spectrum of gradient matrix G:

$$\mathrm{eRank}(G) = \exp\Big(-\sum_i p_i \log p_i\Big), \quad p_i = \frac{\sigma_i(G)}{\sum_j \sigma_j(G)} \quad (2)$$

eRank measures how broadly gradient energy is distributed across singular directions. Higher effective rank implies that gradient energy is distributed across more independent update directions, increasing the dimensionality of the optimisation search space available at each step. We report the ratio $\mathrm{eRank}(G_{\mathrm{FR}})/\mathrm{eRank}(G_{\mathrm{LR}})$ to quantify FR's relative gradient richness.

Gradient starvation. We define *gradient starvation* as the persistent restriction of gradient updates to a low-dimensional update subspace relative to FR training. Operationally, we identify starvation when the ratio $\mathrm{eRank}(G_{\mathrm{FR}})/\mathrm{eRank}(G_{\mathrm{LR}})$ remains substantially above 1 throughout training, indicating that the optimiser cannot compensate for the rank constraint regardless of how many steps are taken. Under starvation, entire update directions available to FR training are absent from the LR update subspace.

Stabilisation mechanism. A run stabilises at checkpoint t_s when self-CKA first exceeds $\tau = 0.995$ and remains above it, indicating that representations are no longer changing meaningfully between updates. We classify the stabilisation mechanism using init-CKA at t_s:

- **Convergence** ($\text{init-CKA}(t_s) \geq 0.65$): representations remain close to initialisation at stabilisation, consistent with a solution found early in the optimisation trajectory.
- **Exhaustion** ($\text{init-CKA}(t_s) < 0.65$): representations have undergone substantial drift from initialisation before stabilising, consistent with the model having explored available update directions without reaching a high-quality solution.

The threshold $\tau_{\text{init}} = 0.65$ is motivated empirically and its sensitivity is discussed in Appendix F.

3 Experimental Setup

3.1 Models and Domains

Language (primary). We pretrain GPT-2 (Radford et al., 2019) on the BabyLM Strict-Small corpus (Hu et al., 2024): 12 transformer layers, 12 attention heads, embedding dimension 768, context length 128. Training uses a 50M-token budget (five passes) with AdamW (Loshchilov and Hutter, 2019), learning rate 10^{-4}, batch size 64, and cosine decay. Checkpoints are saved at Checkpoints are saved at {8K, 32K, 131K, 524K, 1M, 2M, 5M, 10M, 20M, 30M, 40M, 50M} tokens. We evaluate on held-out validation perplexity. All experiments use BF16 precision on NVIDIA A100 GPUs; full hyperparameter tables are provided in Appendix H.

Vision (cross-domain check). We pretrain a ViT (Dosovitskiy et al., 2021) (patch size 4, embed dim 192, depth 9, heads 3) on CIFAR-10 for 20

epochs with AdamW (lr 10^{-3}, cosine decay), saving one checkpoint per epoch. We evaluate on top-1 validation accuracy. Vision results serve as a cross-domain check on the generality of language findings, not as an equal headline contribution.

Why BabyLM. This small-scale controlled setting is a deliberate choice: dense checkpointing is practical, rank sweeps across six values are feasible without distributed-training confounds, and the fixed data budget isolates rank as the causal variable. We hypothesise that the observed mechanisms are not unique to small-scale settings, though validating this at frontier scale remains future work. We do not claim the specific performance numbers generalise; we claim the mechanistic account motivates hypotheses testable at larger scale (Warstadt et al., 2023; Diehl Martinez et al., 2024).

3.2 Low-Rank Parametrisation

For each LR run, all linear projection matrices (query, key, value, output, and both feed-forward projections) are replaced with factorised matrices $W = BA$ at rank r. We sweep $r \in \{8, 16, 32, 64, 128\}$ for vision and $r \in \{8, 16, 32, 64, 128, 256\}$ for language, alongside a full-rank (FR) baseline. All hyperparameters are held constant across FR and LR runs.

Statistical reliability. We train two independently trained seeds per condition, yielding 14 language runs (7 conditions $\times$ 2 seeds) and 12 vision runs (6 conditions $\times$ 2 seeds). Because the paper studies large-effect-size mechanistic trends rather than leaderboard-scale improvements, qualitative reproducibility across seeds is more critical than precise mean estimation. For all major findings — the pre-existing split, gradient eRank ratio trajectory, and stabilisation mechanism classification — qualitative trends are consistent across seeds. Quantitative values are reported as means; where variance is material it is reported as $\pm$ SD. Seed-level results are tabulated in Appendix I. All experiments use AdamW; studying optimizer dependence on gradient starvation is left to future work.

3.3 Probing Pipeline

After training, each checkpoint is probed with a fixed held-out batch of 512 examples drawn from the validation set and held constant across all checkpoints and models, ensuring all measurements are directly comparable. For language, activations are the hidden state of the last non-padding token; for vision, the `[CLS]` token. Layer-wise activations are collected from each transformer block before the residual addition. Cross-run CKA is computed against the corresponding layer of the same fixed FR reference model using the same held-out batch. Layer-wise plots corroborating mean-CKA trends are provided in Appendix G.

Gradient eRank is estimated via a separate forward-backward pass on the same held-out batch. For each weight matrix, the gradient is treated in its native 2D matrix shape; higher-order tensors (e.g., convolutional kernels) are reshaped to their corresponding matrix form before SVD. Economy SVD is applied and eRank computed via Equation (2); reported values are averaged across transformer layers.

3.4 Interventions

Spectral correction at t^*. At the first saved checkpoint t^*, we apply *alpha-flattening*: the singular values of each weight matrix in the LR model are rescaled to a uniform distribution — a high-entropy singular value distribution resembling the comparatively isotropic spectra of randomly initialised FR weights — while singular vectors are left unchanged. This provides intervention evidence for the role of spectral geometry while controlling for other accumulated training differences. We additionally test a *drift-scaled* variant (each matrix scaled by its measured init-CKA drift) and a *combined* variant (both corrections applied simultaneously).

Dynamic spectral normalisation (DSN). To test whether sustained spectral correction can maintain alignment during training, we apply three DSN modes every 1,000 steps, immediately before the optimiser update. *Static*: singular values are rescaled to a fixed uniform target, with Frobenius norm preserved. *Dynamic*: singular values are rescaled to match the distribution of the temporally matched FR checkpoint, obtained from a parallel FR run saved at identical intervals. *Flat*: a uniform singular value distribution is enforced at every step, preserving Frobenius norm. All modes operate on weight matrices directly and do not modify optimiser states.

Layer reinitialisation (k-sweep). To assess post-hoc recoverability, we reinitialise the first k transformer blocks of a trained LR model to full-rank random weights and fine-tune, sweeping $k \in \{0, 1, 2, 3, 4, K\}$ where K denotes all blocks.

Each condition is evaluated under a *short* budget (matched to original training length) and a *convergence* budget (run until validation metric plateaus).

3.5 Normalised Performance Gap

We report separate gap measures for each metric direction. For accuracy (higher is better):

$$\delta_r^{\text{acc}} = \frac{m_{\text{FR}} - m_{\text{LR}}}{m_{\text{FR}}}, \tag{3}$$

and for perplexity (lower is better):

$$\delta_r^{\text{ppl}} = \frac{m_{\text{LR}} - m_{\text{FR}}}{m_{\text{FR}}}. \tag{4}$$

In both cases $\delta_r > 0$ indicates LR underperforms FR and $\delta_r = 0$ indicates parity, making comparisons across metrics interpretable without directional ambiguity.

4 Results

We organise results around six aims. Language findings are reported first throughout; vision findings serve as cross-domain comparisons. Table 1 summarises headline numbers; seed-level variance is reported in Appendix I and is small relative to the effects described here.

Core result. At 8,192 tokens — approximately one batch — LR models at all ranks are already separated from FR in representation space. For LR_r8, cross-run CKA against FR is 0.046 ± 0.003, a 66% reduction relative to the FR–FR baseline (0.136 ± 0.008). The gap changes by less than 0.001 between the first and second checkpoints, indicating the split is established prior to substantial optimisation rather than emerging gradually (Figure 1a). Vision supports this: at epoch 0, before any gradient update, FR–LR CKA for LR_r8 (0.084 ± 0.005) is already far below the FR–FR baseline (0.736 ± 0.012).

Supporting result: rank ordering. Higher rank produces less divergence from FR at every checkpoint in both domains, with monotonic rank ordering stable throughout training. In this setting, LR_r256 reaches near-parity with FR–FR by 50M tokens, while LR_r128 remains below. Exact values are in Table 1.

Domain difference. Normalising FR–FR CKA to t=0, vision FR seeds converge over training ($1.07\times$) while language FR seeds diverge ($0.69\times$) — two independently trained FR models become less similar in language but more similar in vision. This domain-level difference motivates the investigation of whether the split is associated with persistent optimisation differences, addressed in §4.2.

4.1 Aim 2: Language Pretraining Stabilises by Exhaustion

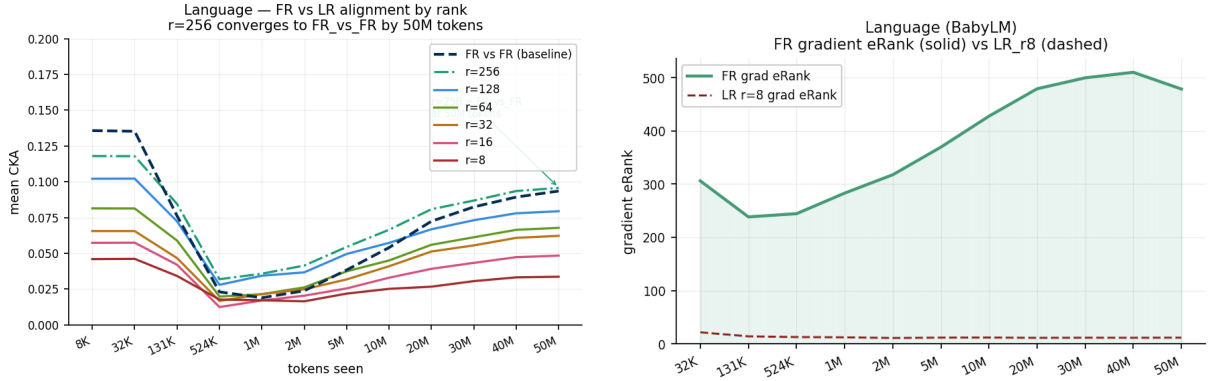
Observation. Under our operational definition (§2), every language run — FR and all LR ranks — stabilises in the exhaustion regime. Mean init-CKA at stabilisation ranges from 0.44 (FR) to 0.67 (LR_r256), all below the 0.65 convergence threshold. In vision, the mechanism is rank-dependent: LR_r128 stabilises by convergence (mean init-CKA $= 0.68$, 100% of layers); LR_r8 and LR_r16 stabilise by exhaustion (≈ 0.13, 100% of layers).

Interpretation. The exhaustion classification for the FR language model itself — not only for LR runs — supports the view that exhaustion-type stabilisation is a property of language pretraining at this scale. Higher rank stabilises earlier in both domains (LR_r256 at $\sim$15–20M tokens; LR_r8 at $\sim$40M tokens in language), which accords with the interpretation that higher-rank models traverse productive gradient directions more rapidly. Alternative explanations are discussed in Appendix F. Together, these observations raise the question of whether LR and FR models differ systematically in gradient dimensionality throughout training, addressed next.

4.2 Aim 3: Evidence for Gradient Starvation

Observation. Figure 2 shows the gradient eRank ratio $\text{eRank}(G_{\text{FR}})/\text{eRank}(G_{\text{LR}})$ across training. In language, the ratio grows substantially throughout training at all ranks: for LR_r8 it rises from 14.3 to 42.3 over the full training run, driven by FR's gradient eRank expanding from 306 to 510 while LR's remains near its structural ceiling. In vision, the ratio is stable across all epochs (CV 2–7% for $r \leq 64$), established at the first update and not closing. Doubling rank approximately halves the final ratio in both domains; exact values are in Table 1.

Interpretation. These observations are consistent with gradient starvation as defined in §2: LR models operate in a persistently reduced update subspace relative to FR, and in language this deficit widens as training progresses. In language, this means the later into training, the further LR falls



(a) Cross-run CKA between FR and LR models across all token checkpoints (BabyLM Strict-Small). Dashed line: FR–FR baseline. All LR ranks fall below FR–FR at 8K tokens; the gap changes by < 0.001 between the first two checkpoints, making a transient initialisation-only effect unlikely. FR–FR CKA decreases over training (language FR seeds diverge; vision FR seeds converge — Appendix A).

(b) Absolute gradient eRank of FR (solid) and LR_r8 (dashed) over training. In language, FR gradient eRank expands from 306 to 510 while LR_r8 remains near its structural ceiling. The growing gap in language underlies the escalating gradient starvation reported in Figure 2 and Appendix C.

Figure 1: **(a) The FR–LR split is established before meaningful learning (Aim 1). (b) FR gradient dimensionality expands in language while LR's remains bounded (Aim 3).** Together these panels motivate the two-phase mechanistic account: the split exists structurally from initialisation and gradient starvation widens it throughout training. Full eRank ratio trajectories in Appendix C.

behind FR's gradient richness — an escalating constraint rather than a fixed one. We cannot rule out that correlated factors contribute to the observed ratios; see §5. This motivates testing whether modifying spectral geometry at initialisation can counteract the divergence, addressed in §4.3.

4.3 Aim 4: Spectral Geometry Drives the Early Split

O4.1 — Intervention at t^*. Alpha-flattening at t^* raises cross-run CKA from 0.082 to 0.318 (+288%) and reduces validation perplexity from 490 to 188 (−62%) in language. Drift-scaling alone provides minimal CKA improvement; the combined variant underperforms alpha-flat alone, supporting the view that spectral geometry rather than drift magnitude is the operative factor. Vision results are consistent: CKA rises by 62% and accuracy improves by 13.5 pp (Table 2). These results provide intervention evidence for the role of spectral geometry at initialisation in driving the early split.

O4.2 — DSN trajectories. *Observation.* Despite perplexity reductions of 75–77% relative to untreated LR (DSN-static: ppl 117; DSN-dynamic: 115; DSN-flat: 107; LR baseline: 457), no DSN mode improves cross-run CKA above the untreated LR baseline at the end of training. Final CKA values cluster near the untreated LR baseline ($\approx$ 0.033–0.036) across all three DSN modes.

Interpretation. Task performance and representational alignment are dissociated under DSN: the intervention substantially improves learning outcomes without producing more FR-like representations. This supports the view that ongoing gradient starvation prevents sustained alignment regardless of spectral initialisation. Crucially, this dissociation — improved perplexity without improved FR-alignment — demonstrates that optimisation success and representational similarity are separable phenomena, a distinction we return to in §5.

4.4 Aim 5: Performance Gap Mirrors Gradient Starvation

The performance gap scaling (Appendix E, Figure 10) closely mirrors the gradient eRank ratio scaling in §4.2: doubling rank approximately halves both the eRank ratio and the normalised performance gap δ_r. In language, δ_r^{ppl} decreases monotonically from 8.72 (r=8) to 0.28 (r=256), with LR_r256 approaching FR at a level consistent with its near-parity CKA in §4. In vision, δ_r^{acc} decreases from 0.59 (r=8) to 0.24 (r=64) but shows a non-monotonic anomaly at r=128 ($\delta = 0.31$), consistent with overfitting at higher capacity on CIFAR-10 and absent in language. Language gaps are $\sim$15$\times$ larger than vision at equivalent ranks, which aligns with the proportionally larger lan-

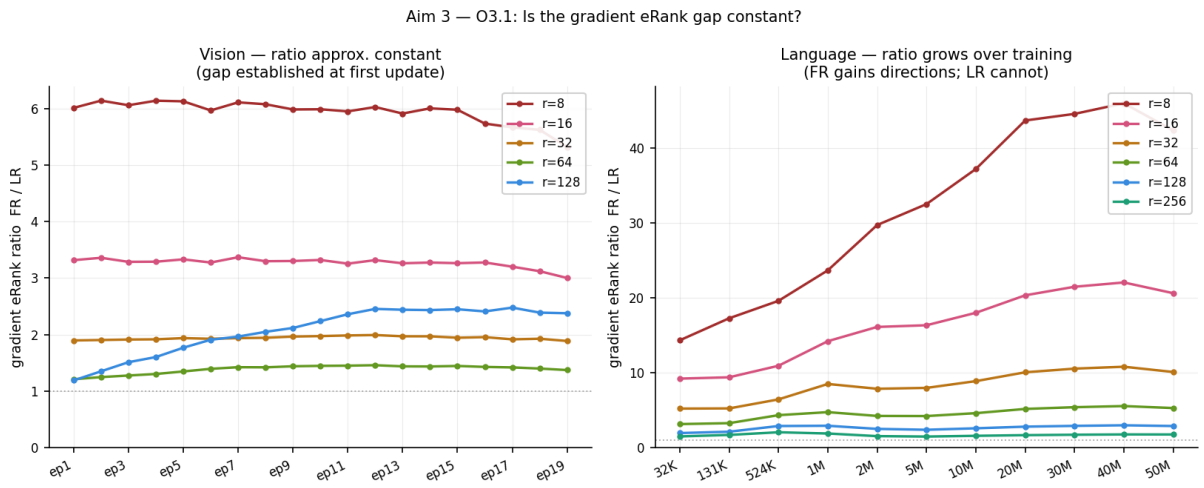


Figure 2: **Gradient starvation sustains and widens the gap throughout training (Aim 3).** Gradient eRank ratio $\mathrm{eRank}(G_{\mathrm{FR}})/\mathrm{eRank}(G_{\mathrm{LR}})$ over training. *Right* (language): the ratio grows substantially throughout training as FR continues to expand its effective gradient dimensionality while LR's remains bounded by the rank constraint. Gradient starvation is an escalating constraint in language. *Left* (vision): the ratio is stable across all epochs (CV 2–7% for $r \leq 64$), established at the first update and not closing thereafter. Gradient starvation is a fixed constraint in vision. Supporting absolute eRank and rank-scaling analyses in Appendix C.

guage eRank ratios in §4.2 and with the theoretical lower bound of Godey et al. (2024). Full results are in Table 1 and Appendix E.

4.5 Aim 6: The Gap is Only Partially Recoverable

Full fine-tuning. Full fine-tuning of a trained LR_r8 model closes 10.7% of the vision gap and 42.0% of the language gap under a convergence budget. Both domains leave substantial gaps unrecovered despite extended optimisation (Table 1 and Appendix E).

Layer reinitialisation. Figure 11 shows recovery as a function of k reinitialised blocks. In *language*, recovery increases monotonically: k=2 closes 31.2%; k=4 closes 54.6%; all blocks (k=11) closes 73.0%. In *vision*, recovery is non-monotonic: k=1 closes 17.2% but k=2 closes only 8.7%, with substantial recovery appearing only at k=9 (all blocks, 49.6%). This all-or-nothing profile in vision, compared with the smooth monotonic profile in language, mirrors the Aim 1 finding that CKA profiles are layer-uniform in language but layer-graded in vision.

Residual gap. Under the most aggressive recovery condition, 51% of the vision gap and 58% of the language gap remain unrecovered. The interpretation of this persistent residual is discussed in §5.

5 Discussion

Our results support a two-phase mechanistic interpretation of low-rank pretraining. Figure 1(a) shows the split is present before meaningful learning — at 8K tokens, all LR ranks are already below the FR–FR baseline, and the gap changes by < 0.001 between the first two checkpoints. Spectral intervention at t^* (Appendix D) raises CKA by 288% in language and reduces perplexity by 62%, supporting the hypothesis that spectral geometry at initialisation is a major contributor. Yet further experiments shows this correction cannot be sustained: all three DSN modes substantially reduce perplexity but leave final CKA near the untreated LR baseline. Figure 1(b) provides evidence for why — FR effective gradient dimensionality continues to expand throughout language training while LR's remains bounded, sustaining the dimensionality gap regardless of spectral correction. *Spectral geometry drives the initial split; gradient starvation sustains it. Correcting either alone is insufficient.*

One of the central findings is the dissociation in optimisation success and representational similarity to FR are separable. DSN achieves substantially better perplexity than untreated LR while final CKA remains unchanged. This indicates that existing performance-based evaluations of low-rank pretraining (Lialin et al., 2024; Zhao et al., 2024; Wei et al., 2024) may systematically underestimate its representational cost. A model can achieve low

Aim	Finding	Language	Vision
1	FR–LR CKA at first checkpoint (LR_r8)	0.046 ± 0.003	0.084 ± 0.005
	FR–FR CKA trend over training	$\downarrow$ 0.69×	$\uparrow$ 1.07×
2	FR stabilisation mechanism	Exhaustion	Mixed
	LR_r8 stabilisation mechanism	Exhaustion	Exhaustion
3	Gradient eRank ratio FR/LR_r8 (final)	42.3	5.3
	Ratio trend over training	Growing	Stable
4	CKA gain from alpha-flat at t^*	+288%	+62%
	DSN final CKA vs LR baseline	No improvement	No improvement
5	Normalised gap δ_r at r=8	8.72	0.59
	Gap scaling with rank	Monotonic	Non-monotonic
6	All-block reinit gap closed	73.0%	49.6%
	Residual gap (best condition)	∼58%	∼51%

Table 1: Headline results across all six aims. Domain differences are informative findings discussed in §5. CKA values are means across seeds and layers; $\pm$ values are standard deviations across seeds.

perplexity while occupying a representational organisation substantially different from a full-rank model — a gap likely relevant for transfer learning (Aghajanyan et al., 2021), probing (Diehl Martinez et al., 2024), and interpretability research. Closing the performance gap and closing the representational gap are distinct targets requiring distinct interventions.

The domain differences (Appendices A–C) further qualify the interpretation. Gradient starvation is a fixed cost in vision but an escalating one in language, where FR continues expanding its effective gradient dimensionality throughout training. Recoverability also differs: language shows smooth monotonic recovery with k reinitialised blocks while vision shows all-or-nothing recovery, mirroring the layer-uniform versus layer-graded CKA profiles in Figure 1(a) and Appendix A. Our findings are therefore strongest as characterisations of language-specific dynamics; the mechanisms may differ at frontier scale, under alternative optimisers, or across architectures and datasets, which we leave to future work. More broadly, our results suggest that efficient pretraining methods should be evaluated not only by downstream performance, but also by the representational and optimisation trajectories they induce.

6 Conclusion

We presented a mechanistic analysis of low-rank pretraining in language models across six objectives. The central finding is a two-phase interpretation: the rank constraint is associated with an immediate representational split at initialisation, and persistent gradient starvation sustains that split throughout training. Correcting either phase alone is insufficient — spectral correction at t^* substantially improves alignment and task performance, but the gains are overwritten by subsequent optimisation dynamics within a few updates.

A key finding is the dissociation between optimisation success and representational alignment: dynamic spectral normalisation reduces perplexity by 75–77% relative to untreated LR while final FR–LR CKA remains unchanged. Closing the performance gap does not imply closing the representational gap — a distinction that performance-only evaluation may fail to detect. If downstream tasks depend on representational structure, as in transfer learning, probing, and interpretability, performance-based evaluation of low-rank pretraining is insufficient. Methods that preserve gradient dimensionality throughout training appear more promising for closing the representational gap than post-hoc spectral correction. More broadly, efficient pretraining methods should be evaluated not only by the behaviours they produce, but also by the representational and optimisation trajectories through which those behaviours emerge. Code and results are released to support future work.[2]

7 Limitations

Scale. All experiments use small models pretrained on a 10M-word corpus. GPT-2 on BabyLM has 85M parameters; our ViT has 4.3M. We hypothesise that the two-phase mechanism — spectral geometry driving the initial split, gradient starvation sustaining it — is not unique to this scale, but we

[2] [anonymised for review].

cannot establish this empirically. At larger scales, the gradient starvation effect may be attenuated if model capacity reduces the relative cost of the rank constraint, or amplified if longer training schedules give the gradient eRank gap more time to widen. Validating these findings at frontier scale remains important future work.

Single optimiser. All experiments use AdamW. The gradient starvation finding may be optimiser-dependent: methods that explicitly expand the gradient subspace, such as second-order optimisers or random gradient projection, might partially counteract the rank constraint. Whether the mechanisms we identify hold under other optimisers is an open question.

Single architecture family. We study standard transformer architectures with dense attention and feed-forward layers. The mechanisms may differ for architectures with different inductive biases — state-space models, mixture-of-experts networks, or linear attention variants — where the role of weight matrix rank is structurally different.

Metrics as proxies. Effective rank is a proxy for gradient dimensionality, not a direct measure of optimisation freedom. CKA measures representational geometry but not functional equivalence: models with low CKA may behave similarly on downstream tasks, and models with high CKA may not. We use both metrics because they are theoretically motivated and practically interpretable, but other analyses — neural tangent kernel comparisons, probing accuracy, or functional similarity metrics — could reveal additional aspects of the FR–LR divide not captured here.

Two seeds. Due to the computational cost of full checkpointed rank sweeps, we use two seeds per condition. This is sufficient to confirm qualitative stability of the large-effect-size trends we report, but does not support precise confidence intervals. Findings that are sensitive to seed — such as the exact stabilisation checkpoint — should be interpreted with appropriate caution.

Causal interpretation. The spectral intervention at t^* provides evidence for the causal role of spectral geometry, but the intervention modifies weight matrices in ways that could affect properties beyond the singular value spectrum, including the alignment of singular vectors across layers. We interpret the results as intervention evidence rather than proof of causation, and note that disentangling the contributions of spectral shape, singular vector geometry, and weight norm remains an open problem.

References

Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.

Richard Diehl Martinez, Pietro Lesci, and Paula Buttery. 2024. Tending towards stability: Convergence challenges in small language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Islam, A. Rupam Mahmood, and Richard S. Sutton. 2024. Loss of plasticity in deep continual learning. In *Nature*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Nathan Godey, Éric de la Clergerie, and Benoît Sagot. 2024. Why do small language models underperform? Studying language model saturation via the softmax bottleneck. *arXiv preprint arXiv:2404.07647*.

Edward J. Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Michael Y. Hu, Aaron Mueller, Ethan Wilcox, Nikita Isac, Isabel Papadimitriou, Daniel Gildea, William Schuler, Naomi Jin, Alex Warstadt, and Ryan Cotterell. 2024. Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at CoNLL 2024*.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International Conference on Machine Learning*.

Jiaxi Li, Lu Yin, Li Shen, Jinjin Xu, Liwu Xu, Tianjin Huang, Wenwu Wang, Shiwei Liu, and Xilu Wang. 2025. Lost: Low-rank and sparse pre-training for large language models. *Preprint*, arXiv:2508.02668.

Vladislav Lialin, Namrata Shivagunde, Sherin Muckatira, and Anna Rumshisky. 2024. ReLoRA: High-rank training through low-rank updates. In *International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, and Will Dabney. 2023. Understanding plasticity in neural networks. In *International Conference on Machine Learning*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. In *OpenAI Blog*.

Olivier Roy and Martin Vetterli. 2007. The effective rank: A measure of effective dimensionality. In *Proceedings of the 15th European Signal Processing Conference*.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*.

Marcus Wei, Maor Ivgi, and Jonathan Berant. 2024. Investigating low-rank training in transformer language models: Efficiency and scaling analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. 2024. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *International Conference on Machine Learning*.

A Aim 1: Supporting Divergence Figures

This section provides the vision replication of the pre-existing split finding and the normalised divergence analyses that support the rank-ordering claim in §4. Figure 3 mirrors Figure 1(a) for vision, confirming that the split is present before any gradient update and that FR seeds converge over training — opposite to the language pattern. Figures 4 and 5 provide the normalised perspective: by expressing FR–LR CKA as a fraction of the FR–FR baseline, the rank ordering becomes directly comparable across domains.

B Aim 2: Stabilisation Analysis

Figure 6 provides the full stabilisation analysis supporting §4.1. Panel (a) shows that higher rank stabilises earlier in both domains, consistent with the interpretation that higher-rank models traverse productive gradient directions more rapidly. Panel (b) shows that all language runs — including FR — stabilise by exhaustion, while vision LR_r128 stabilises by convergence. Panel (c) makes the init-CKA threshold visible directly: convergence runs sit above 0.65, exhaustion runs sit below. Panel (d) reveals the layer gradient in vision: shallow FR layers converge while deep layers exhaust, a pattern absent in language where all layers show uniform exhaustion.

C Aim 3: Supporting Gradient eRank Analyses

Figure 7 supplements the main-text gradient starvation evidence in Figure 2. Panel (a) shows the ratio at the final checkpoint across all ranks: doubling rank approximately halves the ratio in both domains, but language ratios are substantially larger at every rank. Panel (b) separates structural eRank (determined by the rank constraint and largely fixed) from gradient eRank (dynamic, reflecting optimisation). The structural eRank of FR weight matrices decreases slightly over training (spectral consolidation) while gradient eRank is more variable — confirming that the two measures capture qualitatively different properties and that gradient eRank is the more informative diagnostic for the starvation claim.

D Aim 4: Spectral Intervention Details

Figure 8 provides the full O4.1 results: CKA and task-metric comparisons for all four intervention variants at t^* in both domains. Alpha-flattening is the most effective variant in both cases. The combined variant (alpha-flat + drift-scaled) consistently underperforms alpha-flat alone, suggesting that drift scaling adds noise rather than signal — spectral geometry rather than drift magnitude is the operative factor. Figure 9 shows the performance–alignment scatter for both domains, making the

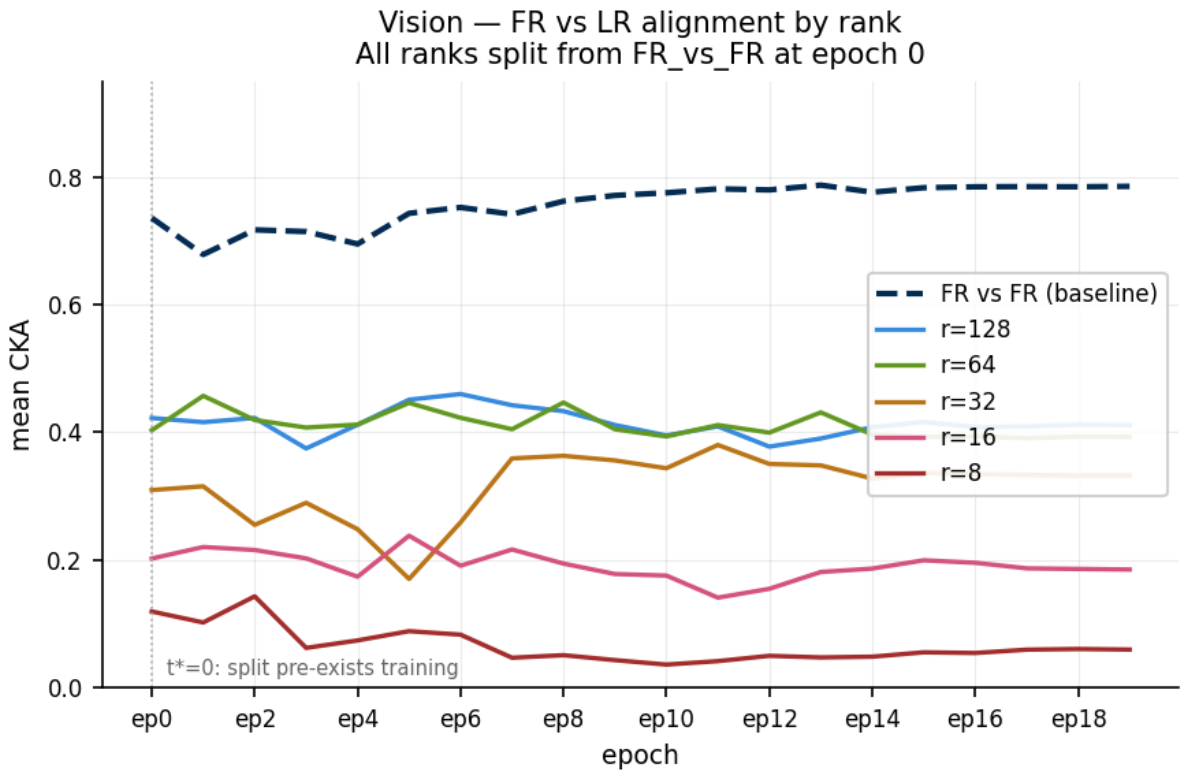


Figure 3: **Vision (CIFAR-10): FR–LR CKA trajectories by rank.** Cross-run CKA between FR and LR models at each rank across all 20 epochs. The split is present at epoch 0, before any gradient update. FR–FR CKA increases over training ($0.736 \rightarrow 0.785$; ratio $1.07\times$): two FR seeds converge in vision, opposite to the language pattern in Figure 1(a). Vision LR_r128 shows a shallower divergence, consistent with its convergence-type stabilisation in Appendix B.

dissociation geometrically visible: DSN conditions occupy the high-performance, low-alignment region regardless of domain or DSN mode.

E Aims 5 and 6: Performance Gap and Recovery

Figure 10 shows the performance gap scaling in full. The key observation is that the gap decreases with rank at approximately the same rate as the gradient eRank ratio in Figure 2: doubling rank approximately halves both. This parallel scaling supports the view that gradient dimensionality is an important mediator of the performance cost, though parallel scaling does not establish a direct causal link.

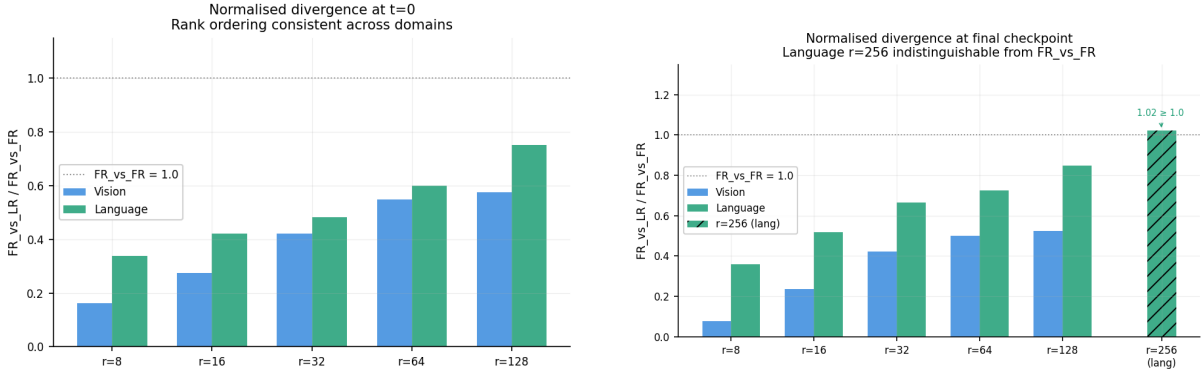
Figure 11 shows the full recoverability results. Panel (a) makes the contrast between language (monotonically increasing recovery) and vision (all-or-nothing recovery) directly visible. Panel (b) shows the residual gap under the most aggressive recovery condition in each domain, confirming that approximately half the gap remains unrecovered regardless of how many layers are reinitialised or how long fine-tuning continues.

F Sensitivity of the Stabilisation Threshold

The convergence/exhaustion classification in §2 uses two thresholds: $\tau_{\text{vel}} = 0.995$ for self-CKA velocity and $\tau_{\text{init}} = 0.65$ for init-CKA at stabilisation. Here we assess the sensitivity of the classification to these choices.

Velocity threshold τ_{vel}. Across $\tau_{\text{vel}} \in \{0.990, 0.995, 0.999\}$, the qualitative ordering of stabilisation points is preserved: higher rank stabilises earlier, and vision LR_r128 stabilises earliest. Absolute checkpoint values shift by at most 1–2 epochs (vision) or 5–10M tokens (language), but no run changes its relative position or mechanism classification. We adopt $\tau_{\text{vel}} = 0.995$ as it best separates runs with clearly stable representations from those still evolving.

Init-CKA threshold τ_{init}. At $\tau_{\text{init}} = 0.55$: all results identical to the main paper. At $\tau_{\text{init}} = 0.75$: vision LR_r64 (two layers) and FR vision (two



(a) Normalised divergence at t=0.

(b) Normalised divergence at final checkpoint.

Figure 4: **FR–LR CKA normalised by FR–FR baseline, both domains.** Values below 1 indicate that FR–LR divergence exceeds natural seed-level variability. Rank ordering is consistent and monotonic across domains at both time points. Language LR_r256 exceeds 1.0 at the final checkpoint (near-parity regime), while LR_r128 remains below 1.0 in both domains.

Domain	Variant	mean CKA	CKA gain	Task metric	Task gain
Language	Baseline	0.082	—	489.7 ppl	—
Language	Drift-scaled	0.104	+27%	478.5 ppl	−2%
Language	Alpha-flat	0.318	+288%	188.2 ppl	−62%
Language	Combined	0.198	+141%	268.2 ppl	−45%
Vision	Baseline	0.318	—	0.223 acc	—
Vision	Drift-scaled	0.362	+14%	0.234 acc	+5%
Vision	Alpha-flat	0.516	+62%	0.253 acc	+13.5pp
Vision	Combined	0.470	+48%	0.249 acc	+12pp

Table 2: Spectral intervention results at t^* (Aim 4, O4.1). CKA gain and task gain are relative to the baseline (untreated LR) in each domain. Alpha-flattening is the most effective variant in both domains; the combined variant underperforms alpha-flat alone.

deep layers) reclassify from convergence to exhaustion; all language runs remain exhaustion. The primary claim — that all language runs stabilise by exhaustion including FR language — holds under all three threshold settings. The convergence classification for vision LR_r128 is somewhat sensitive at the margins and we note this as a qualification on that specific finding.

Alternative interpretation. Exhaustion could reflect a smaller effective step size at stabilisation rather than directional depletion. We assess this by comparing gradient eRank at t_s: vision FR shallow layers (convergence) show 62 ± 4; LR_r8 (exhaustion) show 11 ± 1 — a fivefold difference consistent with the directional interpretation. In language, all runs show similarly low gradient eRank at stabilisation, consistent with the universal exhaustion pattern. A formal decomposition is left to future work.

G Layer-wise CKA Analyses

Mean CKA is reported throughout the main paper. Here we confirm that key findings hold at the individual layer level.

Vision. At epoch 0, all nine ViT layers show FR–LR CKA below the FR–FR baseline, confirming that the pre-existing split holds at every layer individually. Shallow layers (L0–L2) show systematically higher CKA than deep layers (L6–L8) throughout training, consistent with the convergence pattern in shallow FR layers. The layer ordering is stable — no layer crosses another — indicating that the relative difficulty of alignment is fixed early.

Language. All twelve GPT-2 layers are already below the FR–FR baseline at 8K tokens. Unlike vision, language layers show a near-uniform CKA profile across depth (spread < 0.01 at most checkpoints), consistent with the universal exhaustion pattern and the absence of a layer hierarchy. All

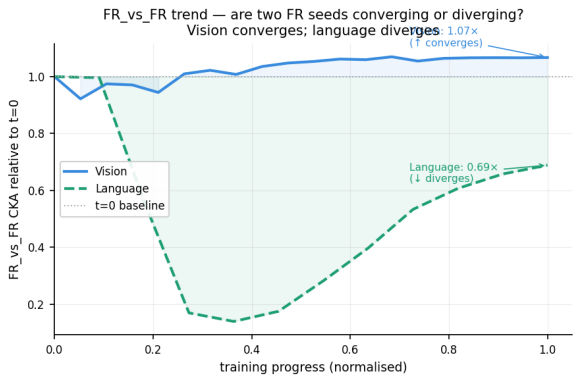


Figure 5: **FR–FR CKA over training, normalised to t=0=1.0.** Both series start at 1.0. Vision FR seeds converge (1.07× by the final epoch); language FR seeds diverge (0.69× by 50M tokens). Fill shading marks convergent (above baseline) and divergent (below baseline) regimes. This domain-level difference is independent of rank and reflects the structure of each domain's optimisation landscape.

layers show a CKA minimum at ∼524K–1M tokens present in both FR–LR and FR–FR comparisons, suggesting it reflects the data distribution rather than the rank constraint. These layer-wise patterns support the use of mean CKA as the primary metric: the mean captures the dominant trend in both domains.

H Compute and Hyperparameter Details

Hardware. All experiments run on NVIDIA A100 80GB GPUs. Language rank sweeps (14 runs × 12 checkpoints) required approximately 280 A100 GPU-hours in total. Vision experiments (12 runs × 20 epochs) required approximately 48 A100 GPU-hours.

Language hyperparameters.

Language hyperparameters.

- Architecture: GPT-2, 12 layers, 12 heads, embed dim 768, context 128, ∼85M parameters (FR)
- Optimiser: AdamW, β_1=0.9, β_2=0.999, ϵ=10^{-8}, weight decay 0.01
- Learning rate: 10^{-4}, cosine decay, 500-step warmup
- Batch size: 64 sequences; token budget: 50M tokens
- Precision: BF16

Vision hyperparameters.

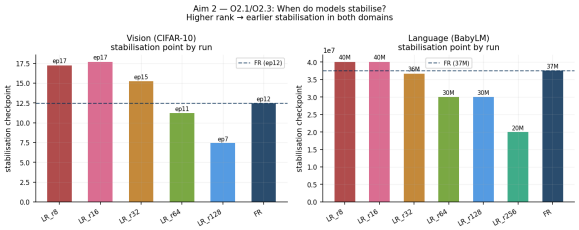
- Architecture: ViT, patch 4, embed dim 192, depth 9, heads 3, ∼4.3M parameters (FR)
- Optimiser: AdamW, β_1=0.9, β_2=0.999, weight decay 0.05
- Learning rate: 10^{-3}, cosine decay, 5-epoch warmup
- Batch size: 128 images; training: 20 epochs on CIFAR-10
- Precision: FP32

I Seed-Level Variance

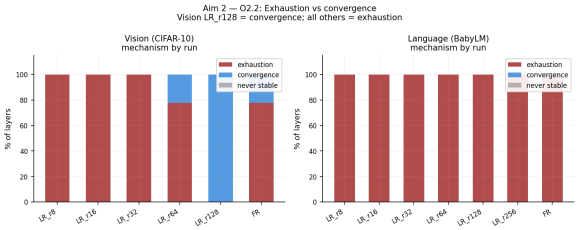
We train two independently seeded runs per condition. Table 3 reports per-seed values for the headline metrics. Variance is consistently small relative to the effect sizes reported in Table 1: the FR–LR CKA split (0.046 vs baseline 0.136) far exceeds the within-condition SD (≤0.005), and the gradient eRank ratio (42.3 for LR_r8) far exceeds any seed-level fluctuation (≤2.1). Qualitative trends — rank ordering, mechanism classification, ratio growth direction — are consistent across both seeds in all conditions.

Metric	Condition	Seed 1	Seed 2
FR–LR CKA (8K)	LR_r8 lang	0.044	0.048
FR–LR CKA (8K)	LR_r8 vis	0.081	0.087
FR–FR CKA (8K)	FR lang	0.132	0.140
FR–FR CKA (8K)	FR vis	0.723	0.748
eRank ratio (final)	LR_r8 lang	40.9	43.7
eRank ratio (final)	LR_r8 vis	5.1	5.5

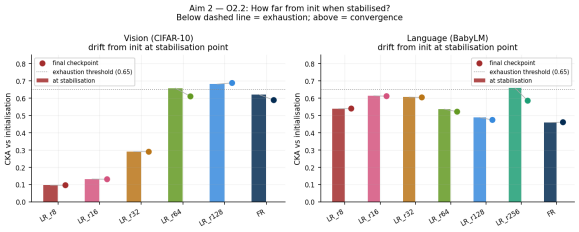
Table 3: Per-seed values for headline metrics. Differences between seeds are small relative to FR–LR effect sizes.



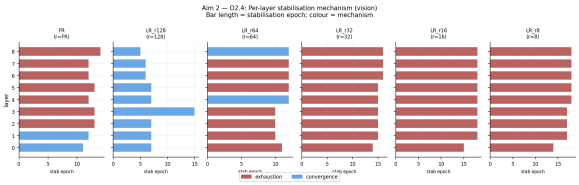
(a) Mean stabilisation checkpoint by run.



(b) % exhaustion / convergence per run.

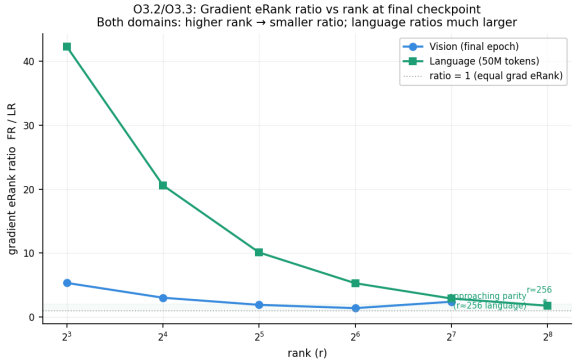


(c) Init-CKA at t_s; dashed line at threshold 0.65.

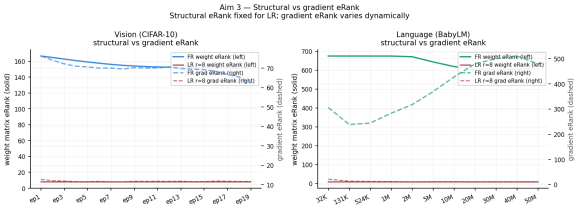


(d) Per-layer mechanism profile (vision).

Figure 6: **Stabilisation analysis, both domains (Aim 2).** (a) Higher rank stabilises earlier in both domains. (b) All language runs (including FR) stabilise by exhaustion; vision LR_r128 stabilises by convergence. (c) Init-CKA at t_s separates the two regimes at threshold 0.65. (d) Vision layer profile: shallow FR layers converge; deep layers exhaust. No such gradient exists in language.

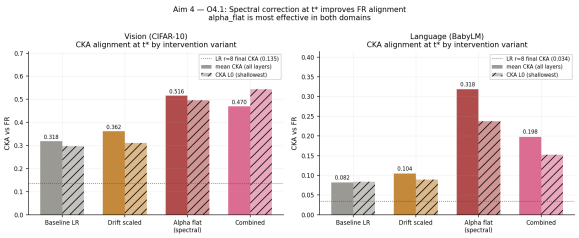


(a) Ratio at final checkpoint vs rank ($\log_2$ scale).

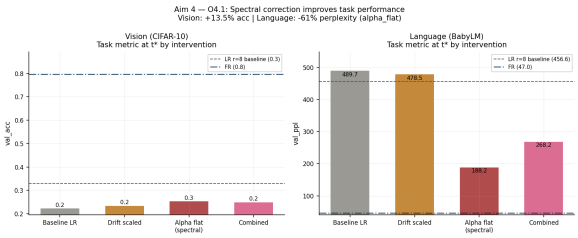


(b) Structural vs gradient eRank over training.

Figure 7: **Supporting gradient eRank analyses (Aim 3).** (a) Doubling rank approximately halves the gradient eRank ratio at the final checkpoint in both domains. Language ratios are $\sim 8\times$ larger than vision at equivalent ranks. (b) Structural (weight matrix) eRank is largely fixed by the rank constraint; gradient eRank is dynamic and not determined by structural eRank. FR weight eRank decreases slightly over training (spectral consolidation); FR gradient eRank expands substantially in language but remains stable in vision.



(a) CKA at t^* by intervention variant.



(b) Task metric at t^* by intervention variant.

Figure 8: **Spectral correction at t^*, both domains (Aim 4, O4.1).** Alpha-flattening is the most effective variant: CKA $+288\%$ (language), $+62\%$ (vision); perplexity -62%; accuracy $+13.5$ pp. The combined variant underperforms alpha-flat alone in both domains, indicating drift scaling adds noise rather than signal. Dashed lines show untreated LR baseline CKA.

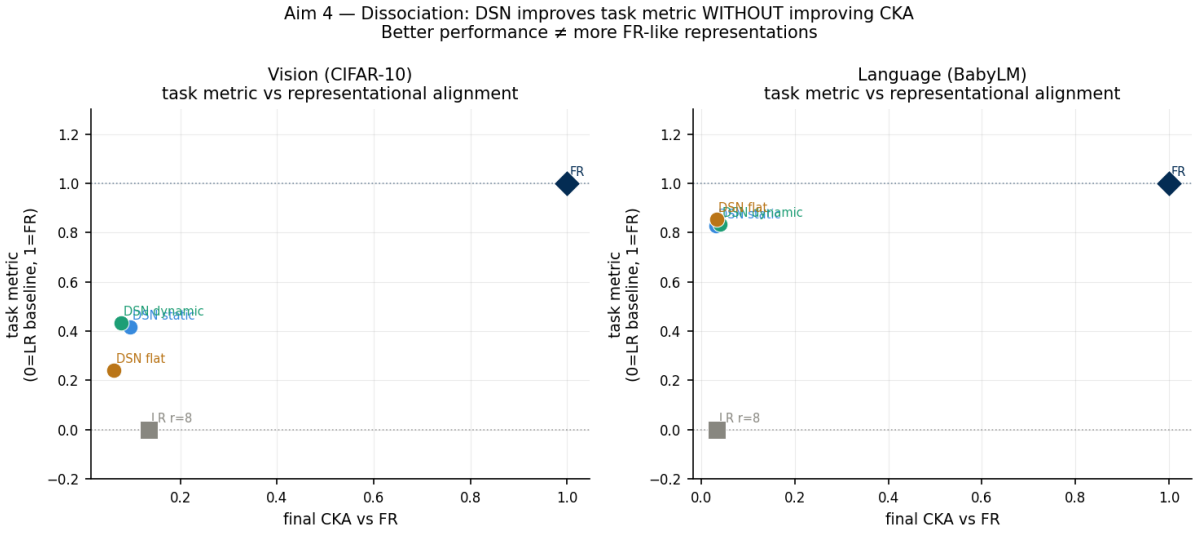
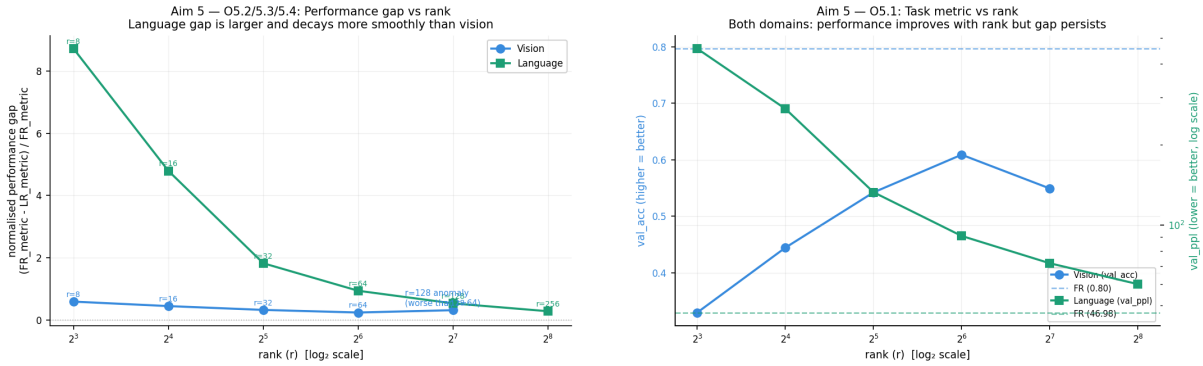


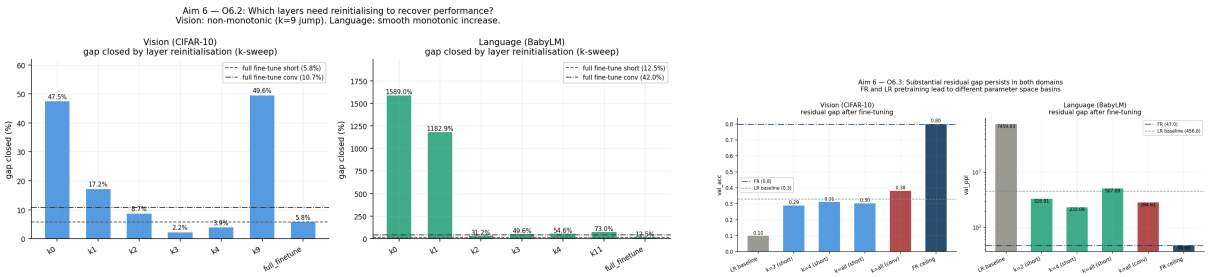
Figure 9: **Task performance vs final CKA, both domains (Aim 4, O4.2).** Each point is a training configuration plotted by final CKA vs FR (x-axis) and normalised task metric (y-axis; 0=LR baseline, 1=FR ceiling). DSN conditions (triangles) cluster in the high-performance, low-alignment region in both domains. FR (diamond) sits at the top right in each panel. The geometric separation confirms that closing the performance gap does not imply closing the representational gap, in both language and vision.



(a) Normalised gap δ_r vs rank ($\log_2$ scale).

(b) Absolute task metric vs rank.

Figure 10: **Performance gap vs rank, both domains (Aim 5).** (a) Language gaps decrease monotonically and are $\sim$15$\times$ larger than vision at equivalent ranks. Vision shows a non-monotonic anomaly at r=128, consistent with overfitting at higher capacity on CIFAR-10. Doubling rank approximately halves the gap in language, mirroring the gradient eRank ratio scaling in Figure 2. (b) Absolute task metrics confirm the pattern; FR reference lines (dashed) show the performance ceiling in each domain.



(a) Gap closed by k-sweep (short budget).

(b) Residual gap under key recovery conditions.

Figure 11: **Recoverability analysis, both domains (Aim 6).** (a) Language: recovery increases monotonically with k; each reinitialised block contributes incrementally, consistent with the layer-uniform CKA profiles in Figure 1(a). Vision: all-or-nothing recovery; substantial improvement only at $k{=}9$ (all blocks), consistent with the layer-graded profiles in Appendix A. Dashed lines show full fine-tune performance under short and convergence budgets. (b) $\sim$51% (vision) and $\sim$58% (language) of the gap remains unrecovered under the most aggressive recovery condition.