Benchmark Shadows: How Data Regimes Shape Parameter Footprints and Generalization

Hongjian Zou[1]* **Yidan Wang**[2]* **Qi Ding**[1] **Yixuan Liao**[1] **Xiaoxin Chen**[1]

Abstract

Large language models can achieve strong benchmark scores without proportional gains in broader capability, but diagnosing when this occurs remains difficult. We study this gap through benchmark shadows: data regimes that may concentrate learning around narrow evaluation-relevant patterns without necessarily improving broader generalization. Within each controlled setting, under fixed architecture, tokenizer, optimizer family, and training budget, we compare a coverage-expanding baseline, a learning-rate control, and early repetition or teacher-rewriting interventions followed by common broad-data training. Repetition causes substantial early degradation whose later mean-score recovery masks slightly higher held-out losses, and rewriting effects depend on model scale and instruction; aggregate spectral proportions detect the learning-rate control and intervention-phase rewriting but do not uniquely characterize behavior, and layer-wise signals vary between baseline reruns. Correlational analyses of open-source multimodal models reveal graded differences in attention-pathway parameters that co-occur with asymmetric reported benchmark profiles, while a prompt de-duplication case study shows that removing surface redundancy can raise benchmark scores despite a lower aggregate spectral proportion. Benchmark scores alone are therefore insufficient to characterize capability; parameter-space diagnostics provide complementary signals about training trajectories but do not by themselves establish generalization or isolate data mechanisms.

1 Introduction

Large language models are commonly evaluated through benchmark suites, and progress is often summarized by score improvements. Yet benchmark gains do not always correspond to proportional gains in broader capability. A model may improve on targeted evaluations while remaining brittle under distribution shift, compositional variation, or long-tail reasoning. This discrepancy raises a central question: when a benchmark score improves, what kind of learning has actually occurred?

A common explanation is contamination or memorization. These effects are important, but they do not exhaust the problem. Even without exact overlap with benchmark examples, a training corpus can be organized around patterns, templates, tasks, or concepts that disproportionately benefit a narrow evaluation distribution. In such cases, the model may develop capabilities that are locally effective but weakly transferable. We refer to this phenomenon as a *benchmark shadow*: an apparent capability gain amplified on targeted evaluations but not matched by comparable improvement in broader generalization.

We study benchmark shadows from a regime-centric perspective. A *data regime* is the higher-level distributional structure induced by sample selection, repetition, rewriting, filtering, and mixture design. Two corpora with similar size can induce different regimes if they differ in semantic coverage, long-tail

*Equal contribution. [1]Vivo AI Lab, Shenzhen, China. [2]Hong Kong University of Science and Technology, Hong Kong, China. Correspondence to: Hongjian Zou (hongjianzou@gmail.com).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

support, or concentration around benchmark-like formats. This perspective complements scaling-centric views: training outcomes are shaped not only by model size, data volume, and optimization, but also by how learning signals are distributed across the corpus. If different regimes reshape weight matrices in distinguishable ways, spectral and rank-based analyses of learned parameters could provide a diagnostic lens that complements benchmark evaluation; whether they do, and at what granularity, is an empirical question we investigate.

This paper asks whether benchmark-shadow-like regimes leave measurable footprints inside model parameters. We combine controlled data interventions with spectral, rank-based, and inter-layer diagnostics, then ask whether related, graded structural differences appear in independently trained multimodal model families where causal identification is not possible but external consistency can be assessed. Our central finding is that controlled data-regime interventions induce distinct structural and behavioral trajectories: mean-score recovery can coexist with small, consistent held-out loss differences, while aggregate spectral summaries capture some training changes without uniquely characterizing behavioral outcomes.

Our main contributions are: (i) we formulate *benchmark shadows* as a regime-level perspective on benchmark-aligned representation formation, distinct from direct contamination or leakage; (ii) we design controlled phase-shift interventions based on redundant repetition and teacher-based rewriting, and characterize their structural and behavioral trajectories, showing that intervention-phase effects and subsequent recovery depend on the model setting and rewriting procedure; (iii) we evaluate aggregate and layer-wise spectral diagnostics against behavioral outcomes and baseline run-to-run variation, finding that the learning-rate control and intervention-phase rewriting exhibit aggregate differences exceeding the observed baseline spread at 20k, whereas small post-recovery differences lack checkpoint-matched replication and localized attention spectral peaks are sensitive to baseline reruns; and (iv) we provide a descriptive external consistency analysis of multimodal models and a prompt de-duplication boundary case showing that removing surface redundancy can improve measured performance despite a lower aggregate spectral proportion.

2 Related Work

Scaling, data, and multimodal training. Scaling laws show that model performance improves predictably with model size, data, and compute under appropriate trade-offs (Sharma and Kaplan [2020], Hoffmann et al. [2022]). However, multimodal systems combine heterogeneous sources, modalities, and supervision formats, making data composition especially important. Prior work on large-scale vision-language training shows that data scale and mixture design both shape visual and cross-modal representations (Jia et al. [2021], Cherti et al. [2023], Aghajanyan et al. [2023]). Recent data-centric work further shows that filtering, domain reweighting, selection, and curation can matter as much as raw scale (Dong et al. [2025]). Our work complements these studies by treating the data regime itself as the object of analysis.

Benchmark effects and narrow specialization. Benchmark performance can be distorted by contamination, benchmark-adjacent curation, and narrow task alignment (Song et al. [2024], Chen et al. [2025]). Related studies show that downstream specialization can improve task-specific metrics while weakening broader transfer (Huang et al. [2024]). We build on this observation but focus on a broader mechanism: benchmark-like data can alter the training regime even without exact benchmark leakage. These concerns are related to Goodhart-style effects and benchmark overfitting, where improvements on measured evaluations need not imply broader gains. Our focus is on controlled data-regime interventions and their behavioral and parameter-space consequences, rather than on exact benchmark leakage or score optimization alone.

Parameter-space diagnostics. Spectral analyses of neural network weights connect empirical spectral structure to optimization, implicit regularization, and generalization (Mahoney and Martin [2019], Martin and Mahoney [2021], Martin and Hinrichs [2025]). Tools such as WeightWatcher provide data-free layer-wise diagnostics based on power-law exponents and related spectral statistics (Martin and Mahoney [2020]). Other work has linked spectral behavior to training dynamics and task difficulty (Xiao et al. [2023], Meng and Yao [2023]). We examine these diagnostics under controlled data-regime interventions and compare them with behavioral outcomes and baseline run-to-run variation, rather than treating spectral summaries as standalone measures of generalization.

3 Benchmark Shadows

We define a *data regime* as the distributional structure of a training corpus beyond its raw size. A regime includes what concepts are covered, how samples are repeated, whether rare patterns are preserved, how supervision formats are distributed, and how strongly the corpus is aligned with downstream evaluations. A *coverage-expanding* regime increases semantic and structural diversity. A *support-concentrated* regime reinforces a narrow subset of patterns or collapses long-tail variation toward canonical forms.

A *benchmark shadow* arises when a support-concentrated regime produces gains on targeted evaluations that are not matched by comparable gains in broader capability. This differs from direct benchmark contamination or data leakage: contamination concerns exact or near-exact overlap between training and test examples, whereas a benchmark shadow can occur without such overlap whenever the training distribution is biased toward patterns over-represented in an evaluation suite. For example, a corpus dominated by document templates may improve document-heavy benchmarks while contributing less to broader reasoning or compositional generalization.

Our hypothesis is that benchmark shadows are visible not only in external behavior but also in parameter-space structure. Different forms of support concentration may induce superficially similar benchmark gains while reflecting fundamentally different changes in the learned model. If a regime encourages broad representation learning, parameter changes should be distributed across layers and subspaces. If a regime encourages shortcut-like adaptation, changes may be localized, spectrally irregular, or weakly recoverable after later training. We examine this hypothesis through controlled interventions and their behavioral and parameter-space outcomes, complemented by descriptive external comparisons.

Diagnostics. We use four complementary diagnostics. The power-law exponent α, fitted to the tail of the eigenvalue distribution of $W^\top W$ above a selected lower cutoff $x_{\min}$ (fitting provenance in Appendix A.8), summarizes weight-matrix spectra; following WeightWatcher-style practice, we track the proportion of matrices with $\alpha \in [2, 6]$ as a coarse aggregate indicator. Effective rank measures singular-mass spread. Delta effective rank measures the effective-rank difference between corresponding matrices in adjacent layers, while change variance measures the entry-wise variance of their weight difference, both within the same checkpoint. These two statistics describe variation across network depth rather than temporal parameter updates. Where relevant, we also report $x_{\min}$ itself and the number of eigenvalues at or above $x_{\min}$ (the fitted-tail eigenvalue count, labeled effective feature number in the analysis records). These metrics are descriptive proxies, not causal mechanisms or standalone measures of generalization; their interpretation requires behavioral evidence and baseline run-to-run calibration.

For external model comparisons, we additionally measure same-layer reference-normalized Frobenius distances, weight correlations, and population variances of weight differences against a common reference model. These metrics characterize parameter-space differences rather than updates from a verified shared initialization.

4 Controlled Data Intervention

We isolate data-regime effects in a controlled text-only setting. Rather than reproducing any specific benchmark-aligned corpus, we model two interventions through which benchmark-oriented construction can distort learning: repeated reinforcement of a narrow pattern family and teacher-based rewriting toward simplified, normalized forms. Architecture, tokenizer, and total budget are fixed across conditions, with one optimization-control baseline that changes only the learning-rate schedule.

4.1 Experimental Setup

Model and training configuration. We train a 0.6B-parameter, 32-layer decoder-only Transformer from scratch with RMSNorm, RoPE, SwiGLU, and grouped-query attention. All conditions share the tokenizer, optimizer family, sequence construction, and approximately 1.2T-token budget. Condition B differs only in its learning-rate schedule. Architecture, optimization, and packing details are provided in Appendix A.2.

Table 1: Proportion of analyzed weight matrices with power-law exponent $\alpha \in [2, 6]$ (%) across training checkpoints for Conditions A–D. Condition B changes the learning-rate schedule, whereas Conditions C and D introduce data-regime interventions.

	2k	**20k**	**40k**	**60k**	**72k**
A: coverage-expanding baseline	60.7	81.7	79.0	79.5	78.6
B: learning-rate control	62.1	57.6	57.6	61.2	61.2
C: repetition regime	58.9	79.9	81.7	80.4	81.3
D: teacher-based rewriting	61.2	77.2	78.1	77.7	77.2

Conditions. Condition A is a deduplicated, coverage-expanding baseline designed to preserve diversity across domains, knowledge types, task families, and instruction formats. It is constructed through a multi-stage data-governance pipeline involving source selection, quality filtering, deduplication, mixture balancing, and targeted augmentation of underrepresented categories. Further details are provided in Appendix A.2. Condition B uses the same training data as A but follows a fast-decay learning-rate schedule that maintains a relatively higher learning rate during the 1.2T-token training run, whereas A, C, and D use cosine decay. Thus, B serves as an optimization-control condition rather than a data intervention. Condition C models *redundant reinforcement*: for the first half of training data $\mathcal{D}_1$, we sample 10% of examples, repeat each selected sample 10 times, and discard the remaining 90%, preserving sample count while sharply reducing support diversity (Marion et al. [2023], Hernandez et al. [2022], Xue et al. [2023]). Condition D models teacher-based rewriting: during the first half of training, source samples are rewritten using an instruction model (Qwen3-30B-A3B-Instruct-2507) according to the prompt in Appendix A.2. Natural-language samples are rewritten using simpler words and sentences, code samples are converted to JavaScript, and exercise samples are left unchanged. The prompt requests preservation of the original meaning and task intent, but semantic equivalence was not independently verified. This approximates aggressive rewriting-based curation or distillation (Algayres et al. [2025]). Because simplification, normalization, and code-language conversion are applied jointly, Condition D represents the combined effects of the rewriting procedure rather than an isolated manipulation of semantic support.

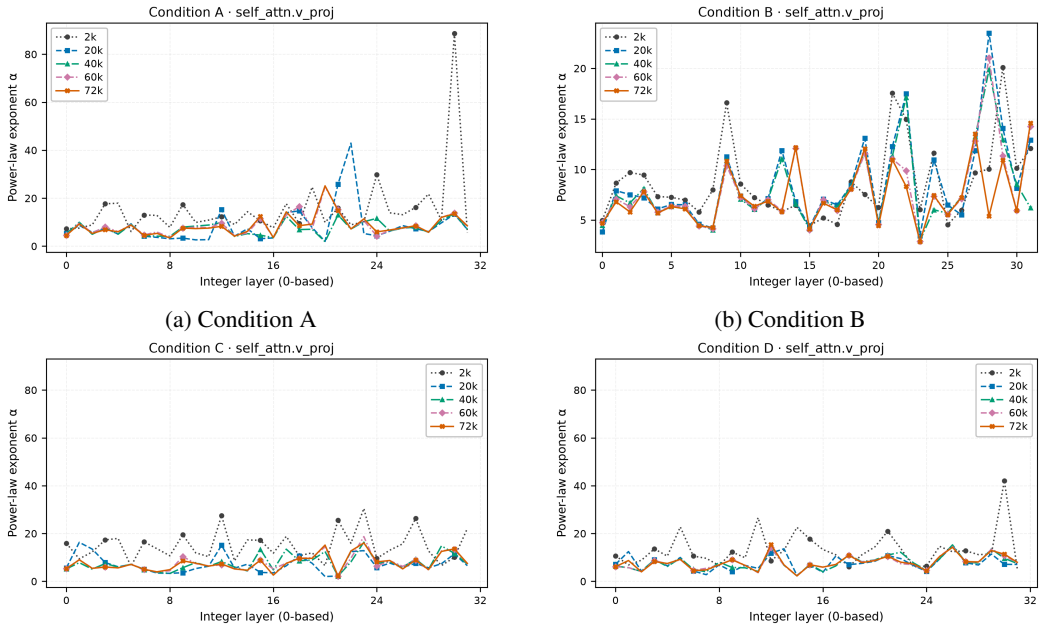
Evaluation protocol. We assess 20 downstream metrics and 33 held-out development-loss datasets (17 in Suite I and 16 in Suite II). Conditions A, C, and D are evaluated at 20k, 60k, and 72k steps; B has available behavioral results at 20k and 60k. Condition B has no reported Suite II results at 20k and only six reported Suite II values at 72k; no complete 72k aggregate is reported. The intervention ends at 36k, so 20k is during intervention and 60k/72k are after common broad-data training. We report unweighted mean and median differences against A and counts of improvements; uncertainty across intervention runs is not established. The evaluation inventory and per-dataset results are in Appendix A.6.

Phase-shift design and scope. Conditions C and D use the same phase-shift curriculum: the first half of training uses a concentrated regime, while the second half reverts to the coverage-expanding baseline. If later diverse training restores the baseline trajectory, the early concentrated regime is largely reversible; if not, the intervention leaves persistent, path-dependent effects. Conditions C and D are controlled abstractions of redundant reinforcement and teacher-based rewriting, not reconstructions of any single real benchmark-aligned corpus.

4.2 Results

Corpus-level effects of rewriting. Across 100,000 paired samples, rewriting shortens the corpus by about 28% while retaining 98.6–99.6% of unique subword types, with nearly unchanged length-controlled lexical diversity. These surface statistics do not establish changes in semantic or task support (Appendix A.5).

Aggregate spectra reflect training settings rather than downstream quality. The learning-rate control provides a clear counterexample to treating aggregate spectral proportions as model-quality scores. Table 1 shows that Condition B has substantially lower proportions of matrices with $\alpha \in [2, 6]$ than A from 20k to 72k, corresponding to differences of approximately 39–54 out of 224 matrices. Yet B achieves the highest mean downstream score at 20k and 60k. The aggregate statistic therefore



(a) Condition A

(b) Condition B

(c) Condition C

(d) Condition D

Figure 1: Layer-wise power-law exponent α in `self_attn.v_proj` for Conditions A–D across training checkpoints (2k, 20k, 40k, 60k, and 72k steps). The profiles show depth-dependent variation across conditions and checkpoints, including localized spectral peaks.

captures large differences associated with the learning-rate schedule without ranking these conditions by downstream performance.

The data interventions produce smaller, checkpoint-dependent spectral differences. At 20k, Condition D has approximately 10 fewer matrices in the interval than A, twice the observed spread of five matrices across three Condition A runs at the same checkpoint. Condition C differs from A by approximately four matrices. Yet at this checkpoint, C performs worse on 18 of 20 downstream metrics and has higher loss on all 33 held-out datasets (Table 2). By 72k, the differences are approximately six matrices for C and three for D. These residual differences are small in magnitude; their reproducibility cannot be assessed using the 20k baseline reference alone (Appendix A.8).

Attention and MLP projections exhibit different diagnostic sensitivity. Aggregate spectral summaries can obscure variation across layers and parameter modules. Figure 1 shows that `self_attn.v_proj` exhibits layer-specific variation across Conditions A–D and training checkpoints, including localized spectral peaks. In contrast, the `mlp.up_proj` spectral profiles share broadly similar depth-wise shapes across Conditions A, C, and D, although their α values differ at some layers (Appendix A.4).

The baseline reruns at 20k reveal a related contrast: isolated spectral peaks in `self_attn.v_proj` vary substantially between the two retrained Condition A models, whereas their `mlp.up_proj` spectral profiles are more similar. Thus, the sensitivity of layer-wise spectral diagnostics depends on the parameter module and the diagnostic used. Localized attention spectral differences require replicate-aware interpretation.

Inter-layer diagnostics characterize depth-wise structural variation. We further examine adjacent-layer effective-rank difference and change variance, the latter computed as the entry-wise variance of weight differences between adjacent layers within the same checkpoint. Both statistics characterize variation across network depth. In the observed runs, change variance exhibits broadly similar depth-wise profile shapes for Conditions A, C, and D, although their raw amplitudes differ. For the available A/C/D records, these amplitude differences largely diminish after normalization by

Table 2: Behavioral differences relative to Condition A at matched training checkpoints. Each entry reports mean difference / median difference / number of wins. Positive downstream-score differences and negative held-out loss differences indicate better performance. Wins count metrics or datasets where the condition outperforms A. Means and medians are computed over 20 downstream metrics, 17 Suite I datasets, and 16 Suite II datasets, respectively. Some downstream metrics share benchmark families and are not statistically independent.

Step	Cond.	Downstream	Suite I	Suite II
20k	C	$-0.040/-0.030/2$	$+0.232/+0.223/0$	$+0.312/+0.320/0$
20k	D	$+0.018/+0.011/15$	$+0.005/+0.008/7$	$+0.003/+0.021/7$
60k	C	$-0.006/-0.002/8$	$+0.014/+0.011/0$	$+0.015/+0.018/5$
60k	D	$+0.018/+0.011/18$	$-0.008/-0.009/15$	$-0.039/-0.037/14$
72k	C	$\approx 0/-0.003/7$	$+0.012/+0.010/0$	$+0.012/+0.015/5$
72k	D	$+0.022/+0.023/15$	$-0.007/-0.008/14$	$-0.025/-0.021/14$

adjacent-layer weight energy. Condition B shows a distinct raw profile under its modified learning-rate schedule (Appendix A.4).

Adjacent-layer effective-rank differences exhibit layer-specific variation across conditions and checkpoints (Figure 8), including localized deviations in C and D. Baseline reruns at 20k also exhibit substantial variation in individual layer-wise differences. These comparisons characterize the evolution and run-to-run sensitivity of inter-layer structure, with localized deviations requiring replicate-aware interpretation.

Behavioral trajectories and recovery. Repetition causes substantial early degradation followed by strong but incomplete behavioral recovery at 0.6B/1.2T. Table 2 shows that Condition C has lower scores on 18 of 20 downstream metrics and higher losses on all 33 held-out datasets at 20k, with mean loss differences of +0.232 and +0.312 on Suites I and II. By 72k, its mean downstream score returns to approximately the baseline level, although only 7 of 20 metrics outperform A. A smaller held-out deficit remains: Suite I mean loss differences are +0.014 at 60k and +0.012 at 72k, with all 17 datasets showing higher loss than A at both checkpoints. Across all 33 held-out datasets, the C–A loss-difference vectors at 60k and 72k correlate at $r = 0.92$, indicating consistency within the observed training trajectory. The Suite II residual is smaller and less consistent across datasets.

Teacher-based rewriting produces a different behavioral trajectory at 0.6B/1.2T. At 20k, Condition D achieves a higher mean downstream score than A (+0.018, with 15 of 20 metrics improving), while its mean held-out loss differences remain small (+0.005 and +0.003 on Suites I and II). The Suite II median difference is +0.021, with only 7 of 16 datasets improving, indicating heterogeneous dataset-level effects. At 60k and 72k, D achieves lower mean losses on both suites and higher mean downstream scores than A. Its downstream mean advantage remains positive across all three checkpoints (+0.018 to +0.022, with 15–18 of 20 metrics improving). At 72k, held-out losses are lower on 14 of 17 Suite I datasets and 14 of 16 Suite II datasets. Among code subsets, JavaScript exhibits the largest loss reduction at all three checkpoints; JavaScript is also the target language of the rewriting intervention.

These observations show contrasting behavioral responses to repetition and teacher-based rewriting, without isolating the contribution of individual rewriting operations.

Cross-scale comparison and recovery. We further examine teacher-based rewriting using 0.6B and 3B models trained from scratch under matched 200B-token budgets. Within each scale, Conditions A, D, and D-remake share the same training configuration, including the initialization procedure, learning-rate schedule, weight decay, batch size, sequence length, and effective token budget. The first approximately 100B tokens are sampled from the corresponding first-half corpora of the original experiment, followed by approximately 100B tokens of common broad-data training. D-remake uses the same teacher as D but employs a rewriting prompt that preserves the original meaning without explicit simplification and retains the original programming language for code. The complete rewriting prompts for D and D-remake are provided in Appendix A.2. We evaluate behavior and spectra at 5k steps, near the end of the intervention phase, and 11k steps, at the end of training (Appendix A.7).

At 0.6B under the 200B-token budget, Condition D shows relatively modest intervention-phase behavioral differences from A. At 5k, its mean downstream score difference is -0.008, with held-out loss differences of +0.009 and +0.041 on Suites I and II. By 11k, the differences become +0.003 in downstream scores and -0.007 and -0.008 in held-out losses. The per-dataset Suite I loss-difference pattern correlates with that of the original 1.2T-token experiment at both the intervention-phase ($r = 0.87$) and final ($r = 0.62$) checkpoints. At 11k, D has lower loss on 14 of 17 Suite I datasets, consistent with the longer-budget observation. The Suite II and downstream mean advantages are smaller at 200B; the final Suite II advantage is sensitive to two English subsets.

At 3B under the same 200B-token budget, teacher-based rewriting produces substantially larger intervention-phase deficits. At 5k, Condition D has higher loss than A on all 33 held-out datasets (mean differences +0.116 and +0.206 on Suites I and II) and lower scores on 18 of 20 downstream metrics (mean difference -0.036). By 11k, the mean loss deficits decrease by approximately 78% and 75%, respectively, although 30 of 33 datasets still have higher loss than A. D-remake shows a similar intervention-phase per-dataset pattern ($r = 0.92$ between the D–A and D-remake–A difference vectors at 5k), but substantially less subsequent recovery. Its mean loss deficits change from +0.110 and +0.166 at 5k to +0.088 and +0.158 at 11k, reductions of approximately 19% and 5%, respectively. The tested 0.6B and 3B settings therefore differ in the initial behavioral response to rewriting, while the two rewriting instructions at 3B show markedly different recovery.

The cross-scale comparisons provide a second example of the limited correspondence between aggregate spectral statistics and behavior. At 11k, the 3B D and D-remake models have identical proportions of matrices with $\alpha \in [2, 6]$ (191/224), despite substantially different held-out loss profiles. The final aggregate difference between D and A is only one matrix at 0.6B and four matrices at 3B. Thus, similar aggregate spectral summaries can accompany different behavioral outcomes. These comparisons involve one run per condition, without checkpoint-matched baseline replication at either scale.

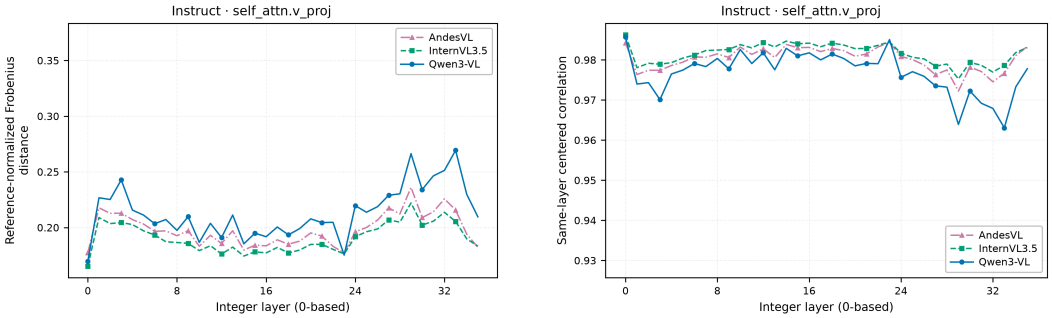
Baseline reproducibility and run-to-run variation. Three independently executed Condition A runs at 20k provide a checkpoint-specific reference for behavioral and spectral variation. The runs include the original baseline, a same-seed rerun, and a different-seed rerun. Pairwise mean absolute differences range from 0.0123 to 0.0204 across the 33 held-out losses and from 0.0134 to 0.0162 across the 20 downstream metrics. Across the three runs, suite-level mean differences span at most 0.005 on Suite I, 0.009 on Suite II, and 0.008 on downstream scores. Condition C's early behavioral degradation is substantially larger than these observed baseline differences, whereas Condition D has small signed mean held-out loss differences with heterogeneous dataset-level effects.

The proportion of matrices with $\alpha \in [2, 6]$ ranges from 79.46% to 81.70% across the three runs, a spread of five out of 224 matrices. Layer-wise comparisons between the two retrained baselines show similar `mlp.up_proj` spectral and change-variance profiles, but substantial variation in individual `self_attn.v_proj` spectral peaks and adjacent-layer effective-rank differences. These results establish diagnostic-dependent run-to-run sensitivity in the tested setting. The observed spread is specific to 0.6B/1.2T at 20k and is not a confidence interval or a reference transferable to later checkpoints or other model scales (Appendix A.8).

Controlled-experiment takeaway. The controlled experiments reveal setting-dependent behavioral responses to data-regime interventions. Aggregate spectral proportions detect some training changes but do not uniquely characterize behavioral effects or recovery. The learning-rate control illustrates that large spectral differences need not imply worse downstream performance. Baseline reruns further show that localized spectral diagnostics require module- and checkpoint-specific calibration; post-recovery differences lack checkpoint-matched replication.

5 External Validation Across Model Families

The controlled experiments characterize regime effects in a simplified setting. We next ask whether related structural differences appear in real systems. This section is necessarily correlational: independently trained models differ in data, architecture, optimization, multimodal design, and alignment procedures. We therefore do not claim that any single observed pattern is caused by one specific training choice. Instead, the goal is external consistency: do structural differences of the kind observed above appear in open-source model families, and do they align with asymmetric benchmark profiles?



(a) Relative parameter change

(b) Weight correlation

Figure 2: Layer-wise comparison of `self_attn.v_proj` with Qwen3-4B-Base for three instruct models. (a) Relative Frobenius change. (b) Pearson correlation of flattened weights. All models remain within a narrow range across depth. Over most layers, Qwen3-VL-4B-Instruct shows the largest relative change and InternVL3.5-4B-Instruct the smallest. Differences are graded rather than categorical.

We compare Qwen3-VL-4B-Instruct, InternVL3.5-4B-Instruct, and AndesVL-4B-Instruct, whose language backbones belong to the Qwen3-4B family. We use Qwen3-4B-Base as a common weight-space reference for matched decoder projections, without assuming identical initialization checkpoints or training histories. AndesVL shows the strongest task-family asymmetry in the reported benchmark profiles.

5.1 Structural Signatures

Aggregate spectra are insufficient. At the model level, the proportion of matrices with $\alpha \in [2, 6]$ depends on matrix inclusion. Over the 252 matched decoder matrices of each instruct model, Qwen3-VL-4B-Instruct, InternVL3.5-4B-Instruct, and AndesVL-4B-Instruct have similar proportions (69.4%, 70.2%, and 70.2%, respectively); their whole-model proportions differ because their non-decoder module inventories differ. These aggregate numbers alone do not explain model behavior. Full model-level α histograms and additional layer-wise spectral comparisons are reported in Appendix A.9; the main text focuses on attention value projections.

Reference-relative parameter differences vary in degree. Layer-wise comparison to the reference checkpoint (Figure 2) shows that all three instruct models differ from the reference in `self_attn.v_proj` within a narrow range. Mean relative Frobenius change is 19.1% for InternVL3.5-4B-Instruct, 19.8% for AndesVL-4B-Instruct, and 21.3% for Qwen3-VL-4B-Instruct, and the mean same-layer correlation with the reference exceeds 0.97 for all three. The ordering holds for most, though not all, layers. The three layer-wise profiles are similar in shape. These comparisons show graded differences in `v_proj` parameter distance, rather than establishing qualitatively distinct forms of decoder reorganization.

Attention value projections were also the module whose spectra varied most across conditions, and across baseline reruns in the controlled experiments of Section 4. This does not imply that any public model followed one of those regimes; it indicates only that `v_proj` is a site at which graded differences between models are visible. Supplementary MLP spectral analyses (Appendix A.9) show comparatively similar `up_proj` α profiles. The thinking variants show the same graded ordering as the instruct variants (Qwen3-VL-4B-Thinking 21.5%, r = 0.976; AndesVL-4B-Thinking 20.1%, r = 0.980; Figure 16).

Modality-specific modules can localize irregularity. The decoder is not the only place where structural footprints appear. Vision-encoder analysis shows that AndesVL exhibits a sharp early-layer peak in the attention query projection `attn.qkv_q`, with a larger fitted-tail eigenvalue count (reported as effective feature number) than the comparison models. The appendix reports this peak (Figure 17) and the corresponding α and fitted lower cutoff $x_{\min}$ profiles (Figure 22). Structural irregularities may appear in either the language decoder or modality-specific submodules.

Table 3: Compact benchmark comparison of representative 4B-scale vision-language models, grouped by task family. AndesVL is competitive on document/perception-heavy tasks but weaker on several reasoning-heavy tasks. Scores are collected from published reports and were not re-evaluated under a shared harness; prompt formats, decoding settings, dataset versions, and evaluation implementations may differ across sources. The table summarizes reported performance profiles and is not a controlled head-to-head comparison.

Family	Benchmark	Qwen3-VL-4B	InternVL3.5-4B	AndesVL-4B
Reasoning	MMMU-Pro	53.2	53.5	37.6
Reasoning	MathVision	51.6	54.4	27.1
Reasoning	DynaMath	65.3	35.7	21.2
Reasoning	LogicVista	53.2	56.4	41.6
Document / perception	AI2D	84.1	82.6	84.5
Document / perception	ChartQA	84.6	86.0	87.8
Document / perception	DocVQA	95.3	92.4	96.0
Document / perception	RealWorldQA	70.9	66.3	72.2

5.2 Capability Asymmetry Across Benchmarks

Structural diagnostics are meaningful only if they relate to external behavior. Table 3 summarizes a compact subset of benchmark results across reasoning-heavy and document/perception-heavy tasks. Qwen3-VL-4B-Instruct and InternVL3.5-4B are stronger on reasoning-intensive evaluations such as MMMU-Pro, MathVision, and DynaMath. AndesVL-4B-Instruct is competitive or stronger on document/perception-oriented evaluations such as ChartQA, DocVQA, AI2D, RealWorldQA, but substantially weaker on several reasoning-heavy tasks. Quantitatively, across the four reasoning-intensive benchmarks, AndesVL trails Qwen3-VL by approximately 24.0 points on average; across four perception-oriented benchmarks with comparable percentage scales (AI2D, ChartQA, DocVQA, and RealWorldQA), it leads by approximately 1.4 points. The reported pattern is therefore asymmetric across task families rather than uniformly weaker performance.

External-validation takeaway. These correlational comparisons reveal graded parameter differences alongside asymmetric reported benchmark profiles, without identifying third-party training regimes or establishing causal mechanisms.

6 Case Study: Prompt Repetition as a Distributional Artifact

The regime framework predicts that not all redundancy has the same effect. To test this boundary, we examine prompt repetition in vision-language training data. Multimodal corpora often duplicate instruction prompts across system messages, user inputs, captions, metadata, OCR-derived text, or templated fields. This creates local structural redundancy, but it does not necessarily narrow semantic support.

Intervention and structural effects. We compare a duplicated-prompt baseline with a de-duplicated variant that removes redundant prompt fields while preserving the underlying image, task target, and image-text supervision. Duplicated segments are verbatim instruction strings or template fragments repeated across fields within the same sample; de-duplication preserves one canonical instance and removes repeated copies. Prompt de-duplication produces selective rather than regime-wide changes: model-level spectra change little and can even move opposite to downstream performance, with the de-duplicated variant showing a lower global $\alpha \in [2, 6]$ proportion despite improving most downstream categories (Appendix A.10). The clearest effects appear in localized attention pathways, and adjacent-layer delta effective rank remains nearly unchanged. This counterexample shows that aggregate spectral summaries are screening signals, not standalone quality measures.

Benchmark effects and takeaway. De-duplication improves four of five evaluation categories (Appendix A.3), with consistent gains on multimodal and instruction-following evaluations while leaving general LLM performance largely unaffected. The Business-LLM average is substantially influenced by a large improvement on keyword extraction (F1), likely reflecting prompt-structure

sensitivity rather than a broad language-capability shift. Surface redundancy can harm model quality and downstream performance, but it does not necessarily induce a regime shift: prompt duplication is a local artifact, whereas the controlled interventions of Section 4 changed the training distribution at the regime level.

7 Discussion

Data regimes shape trajectories, not just scores. Controlled interventions produce distinct structural and behavioral trajectories: after repetition, mean downstream scores return to baseline while small held-out loss differences remain; rewriting effects and recovery depend on model settings and instructions. Correlational external comparisons show graded structural differences alongside asymmetric reported benchmark profiles. Prompt de-duplication improves four of five evaluation categories despite a lower aggregate spectral proportion. Appendix A.1 summarizes the evidence boundaries.

Diagnostics should be multi-level. Aggregate spectral summaries detect some training-condition differences but do not uniquely characterize behavior. Layer-wise diagnostics reveal additional variation, yet localized attention spectral peaks also vary between baseline reruns. Adjacent-layer effective-rank differences and change variance describe variation across network depth rather than temporal parameter updates. Diagnostic reproducibility, discrimination between training conditions, and behavioral relevance are distinct properties that require separate validation.

Evaluation should measure coverage and transfer. Benchmark shadows highlight the limitations of narrow evaluation. Standard benchmarks should be complemented by dataset-level held-out evaluation, robustness and transfer tests, and calibrated parameter diagnostics. These diagnostics provide complementary structural evidence but do not independently establish generalization or isolate data mechanisms.

8 Limitations

This study has several limitations. First, the controlled experiments cover two model scales within one architecture family, with one run per intervention condition. Three baseline runs at 20k provide only an observed variability range, not a confidence interval or a reference for other checkpoints or scales; later held-out loss differences lack matched replication. Second, teacher-based rewriting changes multiple data properties, and D-remake does not isolate individual operations. Neither training budget includes behavioral evaluation exactly at the phase boundary. Third, external comparisons are correlational, involving different model architectures, training procedures, and heterogeneous benchmark sources. Fourth, spectral and rank-based metrics are descriptive proxies, and the adopted $\alpha \in [2, 6]$ interval has not been independently validated for every architecture. Fifth, the external analysis covers a limited set of related model families, restricting generality. Finally, the prompt de-duplication case study examines one artifact and intervention strength, leaving its broader effects unresolved.

9 Conclusion

We introduced benchmark shadows as a regime-level lens for studying the gap between benchmark gains and broader generalization. Controlled interventions reveal distinct structural and behavioral trajectories: mean-score recovery can coexist with residual held-out loss differences, while rewriting effects depend on model settings and instructions. Aggregate spectral summaries detect some training changes but do not uniquely characterize behavioral outcomes; localized diagnostics require baseline run-to-run calibration. Correlational multimodal comparisons and prompt de-duplication provide complementary boundary cases. Overall, parameter-space diagnostics can complement benchmark and dataset-level held-out evaluation, but cannot independently establish generalization or identify data mechanisms.

References

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning*, pages 265–279. PMLR, 2023.

Robin Algayres, Charles-Éric Saint-James, Mahi Luthra, Jiayi Shen, Dongyan Lin, Youssef Benchekroun, Rashel Moritz, Juan Pino, and Emmanuel Dupoux. Longtail-swap: benchmarking language models' abilities on rare words. *arXiv preprint arXiv:2510.04268*, 2025.

Simin Chen, Yiming Chen, Zexin Li, Yifan Jiang, Zhongwei Wan, Yixin He, Dezhi Ran, Tianle Gu, Haizhou Li, Tao Xie, et al. Benchmarking large language models under data contamination: A survey from static to dynamic evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10091–10109, 2025.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023.

Hongyuan Dong, Zijian Kang, Weijie Yin, LiangXiao LiangXiao, ChaoFeng ChaoFeng, and Ran Jiao. Scalable vision language model training via high quality data curation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33272–33293, 2025.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, et al. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*, 2022.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, DDL Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 10, 2022.

Wenke Huang, Jian Liang, Zekun Shi, Didi Zhu, Guancheng Wan, He Li, Bo Du, Dacheng Tao, and Mang Ye. Learn from downstream and be yourself in multimodal large language model fine-tuning. *arXiv preprint arXiv:2411.10928*, 2024.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

Michael Mahoney and Charles Martin. Traditional and heavy tailed self regularization in neural network models. In *International Conference on Machine Learning*, pages 4284–4293. PMLR, 2019.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. When less is more: Investigating data pruning for pretraining llms at scale. *arXiv preprint arXiv:2309.04564*, 2023.

Charles H Martin and Christopher Hinrichs. Setol: A semi-empirical theory of (deep) learning. *arXiv preprint arXiv:2507.17912*, 2025.

Charles H. Martin and Michael W. Mahoney. Weightwatcher: A tool for analyzing neural network weight matrices. `https://github.com/CalculatedContent/WeightWatcher`, 2020.

Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021.

Xuran Meng and Jeff Yao. Impact of classification difficulty on the weight matrices spectra in deep learning and application to early-stopping. *Journal of Machine Learning Research*, 24(28):1–40, 2023. URL `http://jmlr.org/papers/v24/21-1441.html`.

Utkarsh Sharma and Jared Kaplan. A neural scaling law from the dimension of the data manifold. *arXiv preprint arXiv:2004.10802*, 2020.

Dingjie Song, Sicheng Lai, Mingxuan Wang, Shunian Chen, Lichao Sun, and Benyou Wang. Both text and images leaked! a systematic analysis of data contamination in multimodal llm. *arXiv preprint arXiv:2411.03823*, 2024.

Xuanzhe Xiao, Zeng Li, Chuanlong Xie, and Fengwei Zhou. Heavy-tailed regularization of weight matrices in deep neural networks. In *International Conference on Artificial Neural Networks*, pages 236–247. Springer, 2023.

Fuzhao Xue, Yao Fu, Wangchunshu Zhou, Zangwei Zheng, and Yang You. To repeat or not to repeat: Insights from scaling llm under token-crisis. *Advances in Neural Information Processing Systems*, 36:59304–59322, 2023.

A Appendix

A.1 Evidence Scope and Claim Calibration

Table 4: Evidence map for the main claims. The table distinguishes observations from controlled interventions, baseline run-to-run comparisons, correlational external analyses, and a prompt de-duplication case study. Structural diagnostics are treated as descriptive proxies rather than direct evidence of generalization or training mechanisms.

Claim	Evidence	Evidential status	Location
Controlled training regimes produce distinct structural observations.	Aggregate and layer-wise spectral statistics vary across Conditions A–D. Large differences are observed for the learning-rate control, while smaller condition-specific differences are sensitive to baseline run-to-run variation.	Controlled comparisons; spectral reproducibility of intervention effects is not established.	Section 4
Data interventions produce different behavioral trajectories and recovery patterns.	At 0.6B/1.2T, repetition causes substantial early degradation followed by incomplete recovery. At matched 200B budgets, rewriting produces different early effects at 0.6B and 3B; D and D-remake also differ in subsequent recovery.	Controlled, single-run observations; cross-run robustness and underlying mechanisms remain unverified.	Section 4
Aggregate spectral statistics are not standalone measures of model quality.	Condition B exhibits substantially lower aggregate $\alpha \in [2, 6]$ proportions despite favorable downstream scores. At 3B/11k, D and D-remake have identical aggregate proportions but different held-out losses.	Supported descriptively by controlled comparisons; no general predictive relationship is established.	Section 4
Baseline variability limits interpretation of small structural differences.	Three Condition A runs at 20k differ by up to five of 224 matrices in aggregate $\alpha \in [2, 6]$ proportion. Two retrained baselines also show variation in localized attention and rank-based statistics.	Preliminary baseline reference, not a confidence interval or intervention-reproducibility test.	Appendix A.8
Related multimodal models exhibit graded parameter-space differences.	Verified same-layer comparisons with Qwen3-4B-Base reveal nonzero parameter differences and graded attention-projection changes across model families.	Correlational external observation; no causal attribution to a particular training regime.	Section 5
Structural differences can co-occur with asymmetric reported benchmark profiles.	Representative vision-language models show different reasoning-heavy and document/perception benchmark profiles, alongside differences in measured parameter statistics.	Correlational; benchmark scores come from heterogeneous reports, not a shared evaluation harness.	Section 5
Prompt de-duplication has selective structural and behavioral effects.	Removing repeated prompt fields improves several downstream evaluations while producing relatively small aggregate spectral changes.	Single-case intervention; does not establish changes in semantic or task-level support.	Section 6
Parameter-space diagnostics complement behavioral evaluation but do not independently establish benchmark shadows.	Across controlled interventions, cross-scale comparisons, baseline reruns, and external models, aggregate statistics do not uniquely characterize behavioral outcomes or recovery.	Evidence supports diagnostic limitations; benchmark shadows remain a motivating hypothesis rather than a demonstrated mechanism.	Sections 7–8

A.2 Data Construction and Rewriting Details

Model and optimization configuration. The original controlled experiments train a 0.6B-parameter, 32-layer decoder-only Transformer from scratch, with hidden size 1024, feed-forward

dimension 4096, eight attention heads, and grouped-query attention using four query groups. The model uses RMSNorm, RoPE, SwiGLU, a maximum sequence length of 4096, bias-free linear layers, and no attention or hidden-state dropout. The reported initialization is Xavier uniform.

Training uses Adam with $(\beta_1 = 0.9, \beta_2 = 0.95, \epsilon = 10^{-8})$, peak learning rate 3×10^{-4}, cosine decay to 3×10^{-5} unless otherwise specified, global batch size 4096, micro-batch size 4, gradient clipping 1.0, weight decay 0.1, and bf16 precision. Linear warmup spans the first 1.64M samples, followed by learning-rate decay.

Conditions A, C, and D follow cosine learning-rate decay to 3×10^{-5}, whereas Condition B uses a fast-decay schedule that maintains a relatively higher learning rate during training.

Token budget and sequence construction. All conditions use the same tokenizer and a matched training budget of approximately 1.2T tokens. Condition-specific training corpora are prepared separately and packed into 4096-token sequences using a common procedure. The resulting runs therefore have matched training-token exposure, but the number of original source samples is not controlled across conditions. In particular, rewritten texts are shorter on average, so Condition D can contain more source samples within the same token budget.

Condition A: Baseline data construction. The baseline corpus combines heterogeneous Chinese- and English-language sources, including web text, source code, books, academic papers, question answering, and examination materials. The data-governance pipeline includes format normalization, rule-based and model-based quality filtering, and N-gram and MinHash deduplication at multiple granularities. Cross-set decontamination is performed against evaluation benchmarks and held-out development datasets.

A fine-grained content classifier supports mixture balancing across several hundred categories. Underrepresented categories are supplemented through targeted retrieval and importance-weighted resampling. This procedure defines a relatively broad-coverage baseline rather than an independently verified optimum for semantic or task coverage.

Condition C: Repetition. During the first half of training, Condition C selects 10% of the corresponding training examples and repeats each selected example ten times, replacing the remaining 90%. This preserves the nominal example count while substantially increasing repetition and reducing the number of distinct examples. The resulting data are packed using the same procedure and matched training-token budget as the other conditions. During the second half, training returns to the common broad-data regime.

Condition D: Teacher-based rewriting. Condition D uses Qwen3-30B-A3B-Instruct-2507 as the rewriting teacher during the first half of training. The rewriting prompt classifies the input as text, code, or an exercise. Natural-language text is rewritten using simpler words and sentences, code is rewritten in JavaScript, and exercises are requested to remain unchanged.

The prompt requests preservation of the original meaning and task intent, but semantic equivalence is not independently verified. Outputs are subject to validity filtering, including checks for malformed, truncated, or unparsable responses. The accepted rewritten corpus is packed to the same training-token budget as the other conditions.

The intervention jointly changes multiple properties of the training data, including text formulation, sequence length, and code-language distribution. Consequently, the experiment evaluates the implemented rewriting procedure as a whole and does not isolate semantic-support concentration or simplification as an independent causal factor.

The following prompt is used in the teacher-based rewriting procedure for Condition D (teacher-based rewriting). A publicly available instruction model receives each training sample with this instruction, which classifies the content and applies format-specific simplification. English and Chinese variants are used depending on detected language.

English variant.

```
Please rewrite the following content according to these instructions:
1.  First, classify the content as [Text], [Exercise], or [Code].
```

```
2.  If it is [Text], try to use very simple words and sentences to express the
same meaning.
3.  If it is [Code], rewrite it in JavaScript format.
4.  If it is [Exercise], output it exactly as it is without any changes.
5.  Keep formatting symbols in the original text, such as '\n', and do not
change formatting based on them.
6.  Do not add extra content or modify the original data.
7.  Only output the rewritten content; do not mention the classification
result.
8.  Format the output as:  <output> your output </output>.
Content to rewrite:  {text}
```

D-remake rewriting prompt. D-remake uses the same teacher model as Condition D. Its prompt omits the explicit simplification requirement, requests preservation of the original meaning, and retains the original programming language for code. The complete prompt is reproduced below.

English variant.

```
Please rewrite the following content according to these instructions:
1.  First, classify the content as [Text], [Exercise], or [Code].
2.  If it is [Text], rewrite it in your own words, keeping the same meaning,
tone, and level of detail.
3.  If it is [Code], rewrite it in the same programming language as the
original, preserving its behavior; do not translate it into a different
language.
4.  If it is [Exercise], output it exactly as it is without any changes.
5.  Keep formatting symbols in the original text, such as '\n'', and do not
change formatting based on them.
6.  Do not add extra content or modify the original data.
7.  Only output the rewritten content; do not mention the classification
result.
8.  Format the output as:  <output> your output </output>.
Content to rewrite:  {text}
```

Corpus-level effects of rewriting. Corpus-level statistics for 100,000 paired source and rewritten samples are reported in Appendix A.5.

Code availability. The data-construction and diagnostic analysis code, evaluation utilities, and sample outputs are publicly available at `https://github.com/vivo/ScalingLens`. The proprietary training corpora and model checkpoints used in the controlled experiments are not publicly released.

A.3 Prompt De-duplication: Benchmark Scores

A.3.1 General MLLM Benchmarks

Table 5: General-MLLM benchmark scores for the duplicated baseline and de-duplicated variant. Scores are averaged over multiple runs where applicable.

Benchmark	**Baseline**	**De-duplicated**	Δ
MIA-Bench (acc)	0.7558	0.7801	+0.0243
MMMU (pass@4)	0.4756	0.5016	+0.0260
MMBench-dev (acc)	0.7147	0.7240	+0.0093
MMVet	0.4965	0.5412	+0.0448
MMStar	0.5160	0.5451	+0.0291
MathVision (avg@4)	0.1264	0.1381	+0.0117
MathVista (pass@4)	0.6143	0.6213	+0.0070
AI2D-test	0.6803	0.7004	+0.0201
OCRBench	0.7453	0.7507	+0.0053
Average	0.5694	0.5892	+0.0198

A.3.2 General LLM Benchmarks

Table 6: General-LLM benchmark scores for the duplicated baseline and de-duplicated variant.

Benchmark	**Baseline**	**De-duplicated**	Δ
MMLU-Pro (micro acc)	0.1763	0.1850	+0.0087
GPQA-Diamond (pass@4)	0.6330	0.5808	−0.0522
CFBench (psr)	0.3433	0.3467	+0.0033
Average	0.3842	0.3708	−0.0134

A.3.3 Business MLLM Benchmarks

Table 7: Business-MLLM benchmark scores for the duplicated baseline and de-duplicated variant.

Benchmark	**Baseline**	**De-duplicated**	Δ
OCR-GRD (avg score)	0.7185	0.8214	+0.1029
OCR-REC (score)	0.6899	0.7161	+0.0262
Screen-ViVO-Pro (avg score)	0.0724	0.0589	−0.0135
Screen-ViVO-v2.1 (avg score)	0.1870	0.1717	−0.0153
Object Recognition (avg)	0.3589	0.3250	−0.0339
Info Extraction (acc)	0.2041	0.3722	+0.1681
Doc Understanding (attr)	0.2835	0.2994	+0.0159
Doc Understanding (table)	0.4826	0.4728	−0.0099
Image Description (overall)	0.3795	0.3830	+0.0036
Image Grounding (avg score)	0.1214	0.1435	+0.0221
Average	0.3498	0.3764	+0.0266

A.3.4 Business LLM Benchmarks

Table 8: Business-LLM benchmark scores for the duplicated baseline and de-duplicated variant.

Benchmark	**Baseline**	**De-duplicated**	Δ
WritingBench (avg)	0.4813	0.5001	+0.0187
Summary (avg overall)	0.7057	0.7120	+0.0063
Scene NER (overall F1)	0.5539	0.5515	−0.0024
Keywords (F1)	0.1786	0.5896	+0.4110
Average	0.4799	0.5883	+0.1084

A.3.5 Instruction Following Benchmarks

Table 9: Instruction-following benchmark scores for the duplicated baseline and de-duplicated variant.

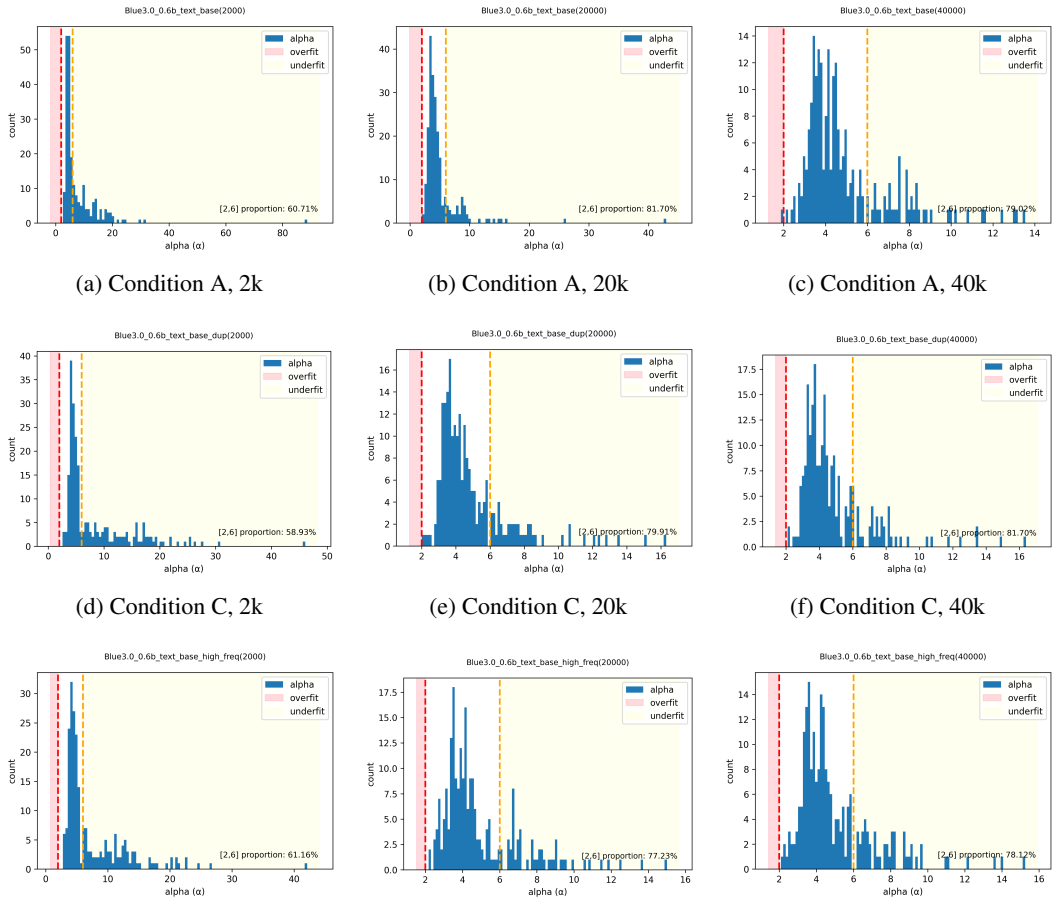
Benchmark	**Baseline**	**De-duplicated**	Δ
IF-v2 (acc)	0.2817	0.2867	+0.0050
CFBench (csr)	0.6067	0.6133	+0.0067
CFBench (isr)	0.2500	0.2567	+0.0067
CFBench (psr)	0.3433	0.3467	+0.0033
IF-Creativity (loose acc)	0.5772	0.6296	+0.0525
IF-Info Extraction (loose acc)	0.7225	0.8197	+0.0973
IF-Summary (loose acc)	0.5666	0.6150	+0.0485
IFEval (inst-level loose)	0.7034	0.6870	−0.0164
IFEval (prompt-level loose)	0.6087	0.5884	−0.0203
MIA-Bench (acc)	0.7558	0.7801	+0.0243
Average	0.5416	0.5623	+0.0208

A.4 Controlled Data Intervention: Additional Figures

The following figures provide supplementary model-level and layer-wise parameter-space diagnostics for the controlled interventions. They illustrate spectral variation across training conditions and checkpoints. As discussed in Section 4, baseline reruns reveal metric-dependent variability, particularly in localized attention-layer statistics. Accordingly, these figures are interpreted as descriptive structural observations rather than evidence of reproducible intervention-specific mechanisms or persistent representational damage.

Layer-wise MLP differences. At 20k, the mean layer-wise `mlp.up_proj` exponents are 4.600 for A, 4.806 for C, and 4.799 for D, despite broadly similar depth-wise profiles. Relative to the observed per-layer range across three baseline records, C exceeds the upper bound at 21 of 32 layers, while D lies outside the range at 22 layers. These are descriptive differences, not evidence of reproducible intervention-specific signatures.

For the 0.6B/1.2T A/C/D records, we normalize adjacent-layer change variance by the combined mean-square weight energy of the two layers. Across 20k–72k, the raw MLP variance differences relative to A range from 11.7% to 28.6%, whereas the corresponding normalized mean differences remain below 0.02% in absolute value. This indicates that the raw amplitude differences are predominantly associated with weight scale.

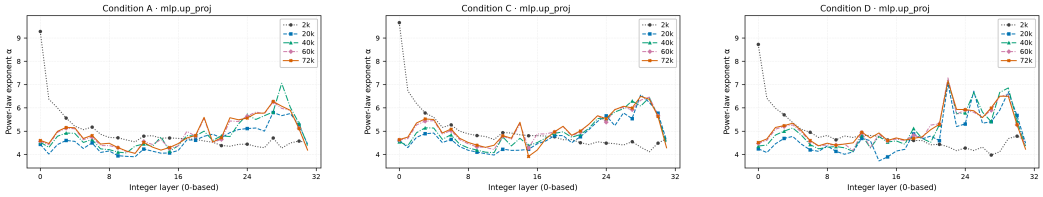


(a) Condition A, 2k (b) Condition A, 20k (c) Condition A, 40k

(d) Condition C, 2k (e) Condition C, 20k (f) Condition C, 40k

(g) Condition D, 2k (h) Condition D, 20k (i) Condition D, 40k

Figure 3: Model-level distributions of power-law exponents α for Conditions A, C, and D at 2k, 20k, and 40k training steps. Rows correspond to the baseline (A), repetition (C), and teacher-based rewriting (D); columns correspond to training checkpoints. The distributions illustrate changes in aggregate spectral structure across conditions and training stages. Their differences are descriptive and do not establish reproducible condition-specific recovery patterns.



(a) Condition A (b) Condition C (c) Condition D

Figure 4: Layer-wise power-law exponent α in `mlp.up_proj` for Conditions A, C, and D across training checkpoints (2k, 20k, 40k, 60k, and 72k steps). The observed profiles are comparatively similar across conditions, in contrast to the more heterogeneous `self_attn.v_proj` profiles shown in Figure 1. This comparison is descriptive and does not establish the reproducibility or functional significance of the attention–MLP differences.

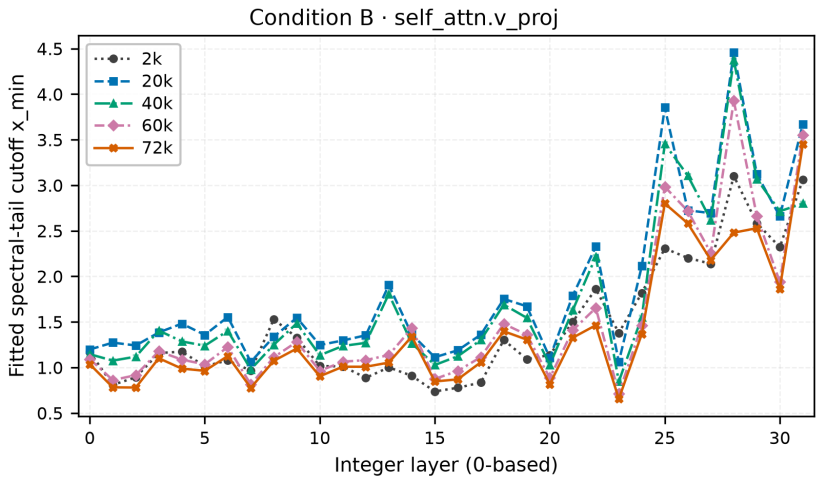
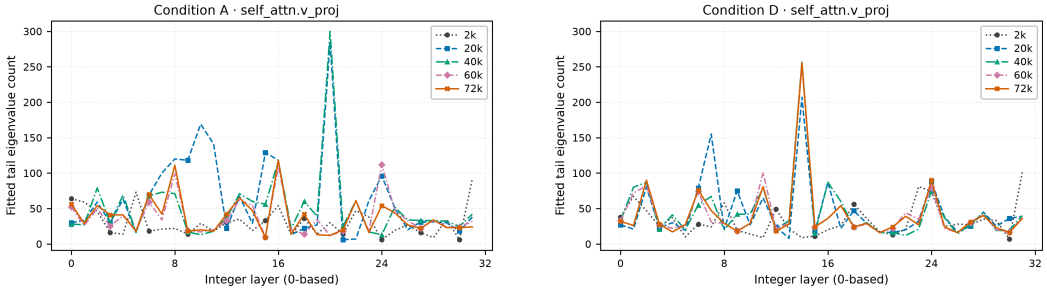
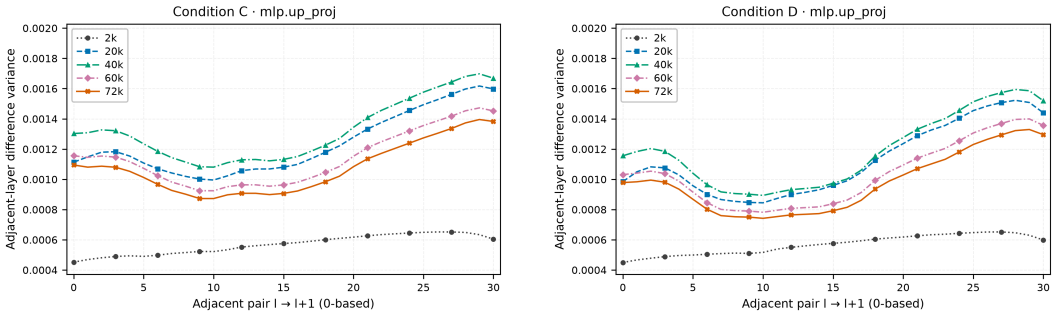


Figure 5: Layer-wise fitted spectral-tail cutoff $x_{\min}$ in `self_attn.v_proj` for Condition B across training checkpoints. The profiles exhibit depth-dependent variation under the modified learning-rate schedule. Here $x_{\min}$ denotes the lower cutoff selected for spectral-tail fitting, not the minimum eigenvalue of the weight matrix. The observed pattern does not directly measure parameter-update dynamics.



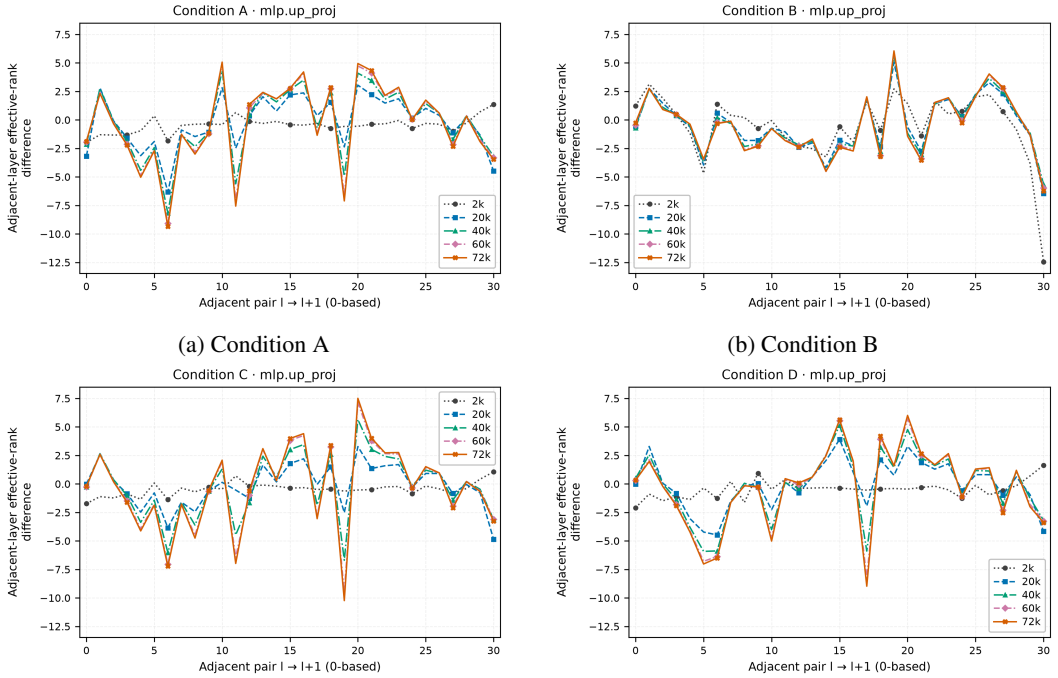
(a) Condition A (b) Condition D

Figure 6: Layer-wise effective feature number in `self_attn.v_proj` for the baseline Condition A and teacher-based rewriting Condition D. The two observed profiles differ at selected layers, including portions of the upper network. These differences describe spectral structure and do not, by themselves, establish changes in functional representational capacity or persistent intervention-specific effects.



(a) Condition C (b) Condition D

Figure 7: Adjacent-layer change variance in `mlp.up_proj` for Conditions C and D across training checkpoints. At each checkpoint, change variance is computed as the entrywise variance of the difference between adjacent-layer weight matrices. The raw variance profiles vary across depth and training checkpoints; their amplitudes are strongly influenced by adjacent-layer weight energy. This statistic does not measure temporal parameter updates or their variance during optimization.



(a) Condition A

(b) Condition B

(c) Condition C

(d) Condition D

Figure 8: Adjacent-layer effective-rank differences in `mlp.up_proj` for Conditions A–D across training checkpoints. Each value is computed as the effective rank of one layer minus that of its preceding layer within the same checkpoint. The profiles show variation across network depth, conditions, and checkpoints.

A.5 Corpus-level Characterization of Condition D

We characterize the distributional changes introduced by Condition D using 100,000 paired samples from the source corpus (A) and their teacher-rewritten counterparts (D). To assess sensitivity to tokenization, we compute corpus-level statistics using two different tokenizers. Table 10 summarizes the results.

Across both tokenizers, rewriting reduces the token count of the paired samples by approximately 28%, while retaining 98.6–99.6% of unique subword types. Length-controlled lexical diversity (MATTR) remains nearly unchanged. Top-1k token coverage and the fitted Zipf exponent increase slightly, whereas Yule's K changes in opposite directions across tokenizers. These results indicate substantial sequence compression without a comparably large shift in the measured subword-level diversity.

These statistics characterize surface token distributions rather than semantic or task-level coverage. They do not establish changes in template, topic, or long-tail support, nor independently verify semantic preservation. The observed structural and behavioral differences therefore cannot be attributed to a particular mechanism of rewriting.

A.6 Detailed Behavioral Results of Controlled Experiments

Evaluation protocol. We evaluate the controlled conditions using 20 downstream metrics and two held-out development-loss suites. The downstream evaluations cover coding (HumanEval), knowledge and reasoning (BBH, C-Eval, GSM8K, C3, GPQA, DROP, MATH, and MMLU), and general-language tasks (CMMLU, AGIEval, Math23K, HellaSwag, ARC, PIQA, BoolQ, and WinoGrande); some benchmarks contribute multiple metrics. For metrics labeled avg4, we report the mean score across four generated responses per item. For pass@4, an item is counted as correct if at least

Table 10: Corpus-level statistics for 100,000 paired source (A) and rewritten (D) samples under two tokenizers. The D/A token ratio and unique-subword retention are reported as ratios; other statistics are shown as A to D.

Metric	Tokenizer 1	Tokenizer 2
D/A token ratio	0.724	0.722
Unique-subword retention	99.6%	98.6%
MATTR	0.482 / 0.480	0.5014 / 0.5016
Top-1k token coverage (%)	67.89 / 68.27	64.77 / 64.99
Zipf exponent	0.9931 / 0.9964	0.9844 / 0.9908
Yule's K	72.90 / 75.28	47.21 / 46.82

one of the four responses is correct. Benchmark-specific few-shot prompts, generation settings, and scoring procedures are held fixed across conditions.

Held-out Suite I contains 17 datasets covering code, mathematics, news, literature, question answering, and world knowledge. Suite II contains 16 additional datasets covering reading comprehension, examinations, high- and low-frequency text, news, and professional content in Chinese and English. The training-data pipeline includes decontamination against evaluation datasets.

Conditions A, C, and D are evaluated at 20k, 60k, and 72k steps; Condition B is evaluated at 20k and 60k where results are available. The intervention ends at 36k steps, placing 20k within the intervention phase and 60k and 72k after the return to common broad data.

We report unweighted mean and median differences relative to Condition A, together with the number of improvements, where available. Higher downstream scores and lower held-out losses are better. These aggregates are descriptive: evaluation metrics and datasets are not necessarily independent, and intervention-level run-to-run uncertainty is not established.

The tables below report the available per-metric downstream scores and per-dataset held-out development losses for the controlled language models. Conditions A, B, C, and D are listed whenever measurements are available at 20k, 60k, and 72k steps for the 0.6B models. Higher downstream scores and lower held-out losses are better. A dash denotes an unreported source value, not a zero. Differences are reported relative to Condition A. Small discrepancies between reported differences and subtraction of displayed values may result from rounding. No uncertainty estimates across independent training runs are available from these tables. Some downstream benchmarks contribute multiple related metrics. For consistent presentation, held-out dataset names use the suffix `-dev`, and downstream metric names include their category prefixes.

Downstream scores (0.6B).

Table 11: 0.6B downstream scores at 20k, 60k, and 72k training steps. $D - A$ and $C - A$ are score differences; higher scores are better.

Step	Metric	A	B	C	D	$D - A$	$C - A$
20000	`Coding/Open-Source_text_humaneval`	0.1146	0.0915	0.0848	0.1707	0.0561	-0.0299
20000	`Knowledge&Reasoning/Open-Source_text_bbhen`	0.3333	0.3723	0.2785	0.3599	0.0266	-0.0548
20000	`Knowledge&Reasoning/Open-Source_text_ceval`	0.2481	0.3507	0.2355	0.3016	0.0535	-0.0126
20000	`Knowledge&Reasoning/Open-Source_text_gsm8ken`	0.4155	0.4306	0.2926	0.4049	-0.0106	-0.1228
20000	`Knowledge&Reasoning/Open-Source_text_c3`	0.3575	0.4452	0.1644	0.3384	-0.0192	-0.1931
20000	`Knowledge&Reasoning/Open-Source_text_gpqa_avg@4`	0.2366	0.2606	0.2589	0.2439	0.0073	0.0223
20000	`Knowledge&Reasoning/Open-Source_text_gpqa_pass@4`	0.6250	0.6786	0.6183	0.6362	0.0112	-0.0067
20000	`General_Text/cmmlu_acc`	0.3158	0.4606	0.2798	0.3475	0.0318	-0.0360
20000	`General_Text/agieval_acc`	0.1016	0.0951	0.0823	0.0926	-0.0090	-0.0193
20000	`General_Text/math23k_cot_acc`	0.1805	0.1474	0.1184	0.2096	0.0291	-0.0622
20000	`Knowledge&Reasoning/Open-Source_text_drop`	0.2474	0.3434	0.2117	0.1751	-0.0723	-0.0357
20000	`Knowledge&Reasoning/Open-Source_text_mathen-fewshot_avg@4`	0.1898	0.1391	0.1361	0.1930	0.0032	-0.0537
20000	`Knowledge&Reasoning/Open-Source_text_mathen-fewshot_pass@4`	0.3506	0.2818	0.2786	0.3552	0.0046	-0.0720

Continued on the next page

Continued from the previous page

Step	Metric	A	B	C	D	$D-A$	$C-A$
20000	Knowledge&Reasoning/Open-Source_text_mmluen	0.2953	0.4263	0.2660	0.3387	0.0433	-0.0293
20000	General_Text/hellaswagen_acc_norm	0.3679	0.5172	0.3527	0.3735	0.0057	-0.0151
20000	General_Text/arc_e_acc	0.3228	0.6102	0.2469	0.3898	0.0670	-0.0758
20000	General_Text/piqa_en_acc_norm	0.6436	0.7051	0.6273	0.6540	0.0103	-0.0163
20000	General_Text/boolq_en	0.5587	0.6364	0.6193	0.6235	0.0648	0.0606
20000	General_Text/winogrande_en_acc_norm	0.5328	0.5620	0.5075	0.5257	-0.0071	-0.0253
20000	General_Text/arc_c_acc	0.2814	0.4102	0.2542	0.3492	0.0678	-0.0271
60000	Coding/Open-Source_text_humaneval	0.1146	0.1159	0.1780	0.1561	0.0415	0.0634
60000	General_Text/agieval_acc	0.1086	0.0933	0.1155	0.1099	0.0013	0.0069
60000	General_Text/arc_c_acc	0.3322	0.4509	0.3559	0.3288	-0.0034	0.0237
60000	General_Text/arc_e_acc	0.4074	0.6561	0.3051	0.4392	0.0317	-0.1023
60000	General_Text/boolq_en	0.6425	0.6835	0.6110	0.6373	-0.0052	-0.0315
60000	General_Text/cmmlu_acc	0.4001	0.4783	0.4029	0.4007	0.0005	0.0028
60000	General_Text/hellaswagen_acc_norm	0.4059	0.5368	0.4045	0.4068	0.0009	-0.0014
60000	General_Text/math23k_cot_acc	0.2257	0.1836	0.2708	0.2658	0.0401	0.0451
60000	General_Text/piqa_en_acc_norm	0.6643	0.7133	0.6730	0.6741	0.0098	0.0087
60000	General_Text/winogrande_en_acc_norm	0.5320	0.5706	0.5375	0.5556	0.0237	0.0055
60000	Knowledge&Reasoning/Open-Source_text_bbhen	0.3963	0.4296	0.3958	0.4364	0.0401	-0.0006
60000	Knowledge&Reasoning/Open-Source_text_c3	0.3904	0.5603	0.3658	0.4370	0.0466	-0.0247
60000	Knowledge&Reasoning/Open-Source_text_ceval	0.2964	0.3804	0.2734	0.3336	0.0372	-0.0230
60000	Knowledge&Reasoning/Open-Source_text_drop	0.2960	0.3712	0.2996	0.3076	0.0115	0.0036
60000	Knowledge&Reasoning/Open-Source_text_gpqa_avg@4	0.2483	0.2578	0.2260	0.2517	0.0033	-0.0223
60000	Knowledge&Reasoning/Open-Source_text_gpqa_pass@4	0.6674	0.6540	0.6116	0.7121	0.0446	-0.0558
60000	Knowledge&Reasoning/Open-Source_text_gsm8ken	0.5610	0.4867	0.5572	0.5701	0.0091	-0.0038
60000	Knowledge&Reasoning/Open-Source_text_mathen-fewshot_avg@4	0.2791	0.1556	0.2751	0.2919	0.0128	-0.0040
60000	Knowledge&Reasoning/Open-Source_text_mathen-fewshot_pass@4	0.4628	0.2946	0.4600	0.4724	0.0096	-0.0028
60000	Knowledge&Reasoning/Open-Source_text_mmluen	0.3777	0.4460	0.3760	0.3807	0.0030	-0.0017
72000	Coding/Open-Source_text_humaneval	0.1122	–	0.1646	0.1988	0.0866	0.0524
72000	General_Text/agieval_acc	0.1197	–	0.1126	0.1174	-0.0023	-0.0071
72000	General_Text/arc_c_acc	0.3153	–	0.3288	0.3424	0.0271	0.0136
72000	General_Text/arc_e_acc	0.4056	–	0.3527	0.4444	0.0388	-0.0529
72000	General_Text/boolq_en	0.6257	–	0.6079	0.6391	0.0135	-0.0177
72000	General_Text/cmmlu_acc	0.4102	–	0.4072	0.3868	-0.0234	-0.0030
72000	General_Text/hellaswagen_acc_norm	0.4075	–	0.4043	0.4070	-0.0005	-0.0032
72000	General_Text/math23k_cot_acc	0.2498	–	0.2648	0.2899	0.0401	0.0150
72000	General_Text/piqa_en_acc_norm	0.6708	–	0.6681	0.6676	-0.0033	-0.0027
72000	General_Text/winogrande_en_acc_norm	0.5264	–	0.5525	0.5533	0.0268	0.0260
72000	Knowledge&Reasoning/Open-Source_text_bbhen	0.4059	–	0.4026	0.4569	0.0509	-0.0033
72000	Knowledge&Reasoning/Open-Source_text_c3	0.3781	–	0.3753	0.4397	0.0616	-0.0027
72000	Knowledge&Reasoning/Open-Source_text_ceval	0.3195	–	0.3105	0.3380	0.0185	-0.0090
72000	Knowledge&Reasoning/Open-Source_text_drop	0.2951	–	0.2941	0.3205	0.0254	-0.0009
72000	Knowledge&Reasoning/Open-Source_text_gpqa_avg@4	0.2478	–	0.2422	0.2500	0.0022	-0.0056
72000	Knowledge&Reasoning/Open-Source_text_gpqa_pass@4	0.6607	–	0.6629	0.6853	0.0246	0.0022
72000	Knowledge&Reasoning/Open-Source_text_gsm8ken	0.5656	–	0.5595	0.5883	0.0227	-0.0061
72000	Knowledge&Reasoning/Open-Source_text_mathen-fewshot_avg@4	0.2831	–	0.2877	0.2998	0.0167	0.0046
72000	Knowledge&Reasoning/Open-Source_text_mathen-fewshot_pass@4	0.4562	–	0.4646	0.4802	0.0240	0.0084
72000	Knowledge&Reasoning/Open-Source_text_mmluen	0.3823	–	0.3740	0.3767	-0.0057	-0.0083

Held-out loss, Suite I (0.6B).

Table 12: 0.6B held-out losses on Suite I at 20k, 60k, and 72k steps. $D - A$ and $C - A$ are loss differences; lower losses are better.

Step	Dataset	A	B	C	D	$D - A$	$C - A$
20000	pure_code-dev	1.4547	1.1714	1.6544	1.4405	-0.0142	0.1996
20000	pure_code_c++-dev	1.2351	1.0322	1.3947	1.2287	-0.0064	0.1596
20000	pure_code_java-dev	1.0594	0.7971	1.2636	1.0565	-0.0029	0.2043
20000	pure_code_javascript-dev	1.3998	1.0830	1.6225	1.3613	-0.0386	0.2227
20000	pure_code_python-dev	1.2373	1.0178	1.4093	1.2454	0.0081	0.1720
20000	pure_en_math-dev	1.4057	1.4781	1.5968	1.4189	0.0133	0.1911
20000	pure_en_news-dev	3.0383	2.6397	3.3145	3.0569	0.0186	0.2762
20000	pure_en_novel-dev	3.4578	2.9175	3.6866	3.4681	0.0103	0.2288
20000	pure_en_qa-dev	2.9982	2.6868	3.1749	3.0234	0.0252	0.1767
20000	pure_en_world_knowledge-dev	2.2598	1.9520	2.4556	2.2621	0.0023	0.1958
20000	pure_hungarian_finals_exam-dev	0.9459	1.1268	1.2035	0.9299	-0.0161	0.2576
20000	pure_skywork_gsm8k-dev	0.9000	0.9516	1.0772	0.9239	0.0239	0.1772
20000	pure_zh_math-dev	1.6271	1.5582	1.8902	1.6162	-0.0108	0.2631
20000	pure_zh_news-dev	3.3327	2.8419	3.6462	3.3528	0.0201	0.3135
20000	pure_zh_novel-dev	3.7368	3.2920	4.0555	3.7223	-0.0145	0.3188
20000	pure_zh_qa-dev	3.9743	3.6753	4.2448	3.9969	0.0226	0.2705
20000	pure_zh_world_knowledge-dev	3.2808	2.8758	3.5985	3.3326	0.0517	0.3177
60000	pure_code-dev	1.35740	1.15015	1.36716	1.34123	-0.0162	0.0098
60000	pure_code_c++-dev	1.15701	1.01863	1.15941	1.14656	-0.0104	0.0024
60000	pure_code_java-dev	0.97642	0.78551	0.99042	0.96914	-0.0073	0.0140
60000	pure_code_javascript-dev	1.28826	1.05697	1.29947	1.26494	-0.0233	0.0112
60000	pure_code_python-dev	1.13935	0.99193	1.15202	1.13306	-0.0063	0.0127
60000	pure_en_math-dev	1.33919	1.44992	1.34352	1.32967	-0.0095	0.0043
60000	pure_en_news-dev	2.90161	2.60917	2.90972	2.89377	-0.0078	0.0081
60000	pure_en_novel-dev	3.31900	2.88697	3.32876	3.31723	-0.0018	0.0098
60000	pure_en_qa-dev	2.88964	2.66139	2.89841	2.88140	-0.0082	0.0088
60000	pure_en_world_knowledge-dev	2.14022	1.92976	2.14713	2.13182	-0.0084	0.0069
60000	pure_hungarian_finals_exam-dev	0.87192	1.08913	0.94451	0.89414	0.0222	0.0726
60000	pure_skywork_gsm8k-dev	0.85922	0.93192	0.87830	0.86229	0.0031	0.0191
60000	pure_zh_math-dev	1.51653	1.51753	1.52727	1.50682	-0.0097	0.0107
60000	pure_zh_news-dev	3.17493	2.81033	3.18965	3.16439	-0.0105	0.0147
60000	pure_zh_novel-dev	3.58442	3.26212	3.59504	3.56319	-0.0212	0.0106
60000	pure_zh_qa-dev	3.82914	3.64469	3.84268	3.81297	-0.0162	0.0135
60000	pure_zh_world_knowledge-dev	3.13348	2.84149	3.14205	3.12456	-0.0089	0.0086
72000	pure_code-dev	1.3508	–	1.3590	1.3339	-0.0170	0.0081
72000	pure_code_c++-dev	1.1527	–	1.1545	1.1434	-0.0094	0.0018
72000	pure_code_java-dev	0.9706	–	0.9807	0.9610	-0.0096	0.0100
72000	pure_code_javascript-dev	1.2819	–	1.2910	1.2560	-0.0259	0.0091
72000	pure_code_python-dev	1.1287	–	1.1407	1.1248	-0.0039	0.0120
72000	pure_en_math-dev	1.3246	–	1.3345	1.3264	0.0018	0.0099
72000	pure_en_news-dev	2.8907	–	2.8978	2.8827	-0.0080	0.0071
72000	pure_en_novel-dev	3.3109	–	3.3159	3.3060	-0.0050	0.0050
72000	pure_en_qa-dev	2.8810	–	2.8895	2.8741	-0.0069	0.0086
72000	pure_en_world_knowledge-dev	2.1346	–	2.1391	2.1231	-0.0114	0.0045
72000	pure_hungarian_finals_exam-dev	0.8730	–	0.9237	0.8969	0.0239	0.0506
72000	pure_skywork_gsm8k-dev	0.8539	–	0.8708	0.8644	0.0105	0.0168
72000	pure_zh_math-dev	1.5091	–	1.5195	1.5015	-0.0076	0.0104
72000	pure_zh_news-dev	3.1629	–	3.1758	3.1527	-0.0102	0.0129
72000	pure_zh_novel-dev	3.5693	–	3.5795	3.5517	-0.0176	0.0101
72000	pure_zh_qa-dev	3.8156	–	3.8288	3.8029	-0.0128	0.0132
72000	pure_zh_world_knowledge-dev	3.1182	–	3.1290	3.1131	-0.0051	0.0107

Held-out loss, Suite II (0.6B).

Table 13: 0.6B held-out losses on Suite II at 20k, 60k, and 72k steps. $D - A$ and $C - A$ are loss differences; lower losses are better.

Step	Dataset	A	B	C	D	$D - A$	$C - A$
20000	Acient_Chinese_reading_2025_GK-dev	4.7517	–	5.2017	4.8085	0.0567	0.4500
20000	English_reading_easy-dev	2.9673	–	3.1243	2.9064	-0.0609	0.1571
20000	GRE_202501-dev	3.4740	–	3.7414	3.5500	0.0760	0.2674
20000	HSK5-dev	3.6955	–	4.1008	3.6428	-0.0527	0.4053
20000	Morden_Chinese_reading_2025_GK-dev	4.2231	–	4.6017	4.2171	-0.0060	0.3785

Continued on the next page

Continued from the previous page

Step	Dataset	A	B	C	D	$D-A$	$C-A$
20000	high_freq_cn-dev	3.9148	–	4.2641	3.9560	0.0412	0.3493
20000	high_freq_cn_sampled-dev	3.9710	–	4.3303	4.0055	0.0345	0.3593
20000	high_freq_en-dev	2.9997	–	3.1770	2.9510	-0.0487	0.1773
20000	history_geography_politics_2025_GK-dev	3.6973	–	4.0537	3.7462	0.0488	0.3564
20000	low_freq_cn-dev	3.8554	–	4.2704	3.8375	-0.0179	0.4150
20000	low_freq_en-dev	3.2939	–	3.5487	3.3259	0.0320	0.2548
20000	low_freq_en_sampled-dev	3.4840	–	3.7670	3.5430	0.0590	0.2830
20000	news_cn-dev	3.4144	–	3.7657	3.4577	0.0433	0.3513
20000	news_en-dev	3.2777	–	3.5058	3.2566	-0.0211	0.2280
20000	professional_cn-dev	3.1703	–	3.4613	3.0222	-0.1481	0.2910
20000	professional_en-dev	2.4762	–	2.7436	2.4863	0.0101	0.2673
60000	Acient_Chinese_reading_2025_GK-dev	4.5347	4.5740	4.5319	4.5223	-0.0124	-0.0029
60000	English_reading_easy-dev	2.8589	3.5292	2.8559	2.8215	-0.0373	-0.0030
60000	GRE_202501-dev	3.2874	3.2211	3.3273	3.2916	0.0041	0.0399
60000	high_freq_cn-dev	3.7333	3.7465	3.7522	3.7150	-0.0183	0.0189
60000	high_freq_cn_sampled-dev	3.8111	3.7743	3.8381	3.7704	-0.0407	0.0270
60000	high_freq_en-dev	2.9083	3.5159	2.8954	2.8586	-0.0497	-0.0129
60000	history_geography_politics_2025_GK-dev	3.5521	3.5255	3.5711	3.5304	-0.0217	0.0189
60000	HSK5-dev	3.5044	3.6634	3.5619	3.4656	-0.0388	0.0575
60000	low_freq_cn-dev	3.6057	3.6838	3.6303	3.5686	-0.0371	0.0246
60000	low_freq_en-dev	3.0872	3.0764	3.1176	3.0679	-0.0193	0.0304
60000	low_freq_en_sampled-dev	3.2772	3.0494	3.3234	3.2776	0.0004	0.0462
60000	Morden_Chinese_reading_2025_GK-dev	4.0279	4.2005	4.0449	3.9816	-0.0464	0.0170
60000	news_cn-dev	3.2554	3.6717	3.2567	3.2308	-0.0247	0.0012
60000	news_en-dev	3.1552	3.4544	3.1458	3.0963	-0.0589	-0.0094
60000	professional_cn-dev	2.8616	3.1368	2.8729	2.7573	-0.1043	0.0113
60000	professional_en-dev	2.2660	2.5196	2.2485	2.1547	-0.1112	-0.0175
72000	Acient_Chinese_reading_2025_GK-dev	4.5197	–	4.5044	4.5111	-0.0086	-0.0153
72000	English_reading_easy-dev	2.8460	–	2.8410	2.8859	0.0399	-0.0050
72000	GRE_202501-dev	3.2838	–	3.3209	3.2798	-0.0040	0.0371
72000	HSK5-dev	3.4839	–	3.5383	3.4504	-0.0334	0.0545
72000	Morden_Chinese_reading_2025_GK-dev	4.0205	–	4.0258	3.9760	-0.0445	0.0052
72000	high_freq_cn-dev	3.7166	3.7330	3.7330	3.6984	-0.0181	0.0164
72000	high_freq_cn_sampled-dev	3.7965	3.7613	3.8155	3.7622	-0.0343	0.0190
72000	high_freq_en-dev	2.9002	3.5566	2.8849	2.9481	0.0478	-0.0154
72000	history_geography_politics_2025_GK-dev	3.5380	–	3.5523	3.5211	-0.0169	0.0143
72000	low_freq_cn-dev	3.5811	3.6704	3.6126	3.5609	-0.0202	0.0315
72000	low_freq_en-dev	3.0816	3.0649	3.1098	3.0572	-0.0245	0.0281
72000	low_freq_en_sampled-dev	3.2776	3.0437	3.3209	3.2698	-0.0078	0.0433
72000	news_cn-dev	3.2405	–	3.2433	3.2195	-0.0210	0.0028
72000	news_en-dev	3.1443	–	3.1300	3.1006	-0.0437	-0.0144
72000	professional_cn-dev	2.8234	–	2.8529	2.7441	-0.0793	0.0294
72000	professional_en-dev	2.2604	–	2.2183	2.1351	-0.1253	-0.0421

A.7 Cross-scale Comparison and Prompt-Variant Analysis

Experimental design. We conduct additional controlled experiments at two model scales, 0.6B and 3B, under matched training budgets of approximately 200B tokens. All models are trained from scratch. Within each scale, Conditions A, D, and D-remake share the same model architecture, initialization procedure, optimizer configuration, learning-rate schedule, weight decay, batch size, sequence construction, and effective training-token budget.

The experiments follow the two-phase design of the main controlled study. During the first phase, approximately 100B tokens are drawn from condition-specific corpora. For Conditions A and D, these data are sampled from the corresponding first-half corpora of the 1.2T-token experiment. D-remake uses the same teacher as D but employs a rewriting prompt that removes the explicit simplification requirement, requests preservation of the original meaning, and retains the original programming language for code. The rewriting procedures and complete prompts for D and D-remake are provided in Appendix A.2. During the second phase, all conditions receive approximately 100B tokens from the common broad-data regime.

The 5k checkpoint lies within the intervention phase, near its end; the 11k checkpoint represents the end of training. The interval from 5k to 11k therefore includes the remaining intervention-phase training as well as the subsequent common-data phase. We do not treat 5k as an exactly measured phase-transition checkpoint.

Table 14: Cross-scale behavioral differences relative to Condition A at 5k and 11k steps. Each cell reports unweighted mean difference / number of improvements. Downstream scores are higher-is-better; held-out losses are lower-is-better. The denominators are 20 downstream metrics, 17 Suite I datasets, and 16 Suite II datasets. D-r denotes D-remake.

Scale	Cond.	Step	Downstream	Suite I	Suite II
0.6B	D	5k	$-0.0076/6$	$+0.0088/6$	$+0.0412/4$
	D	11k	$+0.0029/10$	$-0.0065/14$	$-0.0083/10$
	D-r	5k	$-0.0149/7$	$+0.0119/7$	$-0.0237/9$
	D-r	11k	$-0.0059/9$	$+0.0117/6$	$-0.0100/7$
3B	D	5k	$-0.0362/2$	$+0.1160/0$	$+0.2056/0$
	D	11k	$-0.0225/3$	$+0.0258/2$	$+0.0508/1$
	D-r	5k	$-0.0593/1$	$+0.1095/1$	$+0.1657/0$
	D-r	11k	$-0.0421/2$	$+0.0883/1$	$+0.1582/0$

Both checkpoints are evaluated using the same 20 downstream metrics and 33 held-out loss datasets (17 in Suite I and 16 in Suite II). We report differences relative to the corresponding Condition A at the same scale and checkpoint. Downstream scores are higher-is-better; held-out losses are lower-is-better. Aggregate differences are unweighted across reported metrics or datasets.

Behavioral trajectories. Table 14 summarizes the intervention-phase and final behavioral results.

At 0.6B, Condition D exhibits relatively small intervention-phase differences in Suite I (+0.0088), with a larger mean deficit on Suite II (+0.0412). At 11k, the mean differences become -0.0065 and -0.0083, respectively. Mean downstream performance moves from -0.0076 to +0.0029 relative to A. The final evaluation therefore shows approximately baseline-level downstream performance and small average held-out loss advantages.

At 3B, the intervention-phase response differs substantially. At 5k, Condition D performs worse than A on all 33 held-out loss datasets, with mean differences of +0.1160 on Suite I and +0.2056 on Suite II. Only 2 of 20 downstream metrics improve. At 11k, the corresponding loss deficits decrease to +0.0258 and +0.0508, representing reductions of approximately 78% and 75% in the respective mean gaps. Nevertheless, 30 of 33 held-out datasets still have higher loss than A, and 17 of 20 downstream metrics remain lower.

The 3B D-remake variant exhibits a different recovery pattern. Its initial mean held-out loss deficits are +0.1095 and +0.1657 on Suites I and II, smaller than those of D at 5k. By 11k, however, these deficits remain +0.0883 and +0.1582, corresponding to much smaller reductions of approximately 19% and 5%. Consequently, the two rewriting variants reverse their relative ordering in aggregate held-out loss between the two checkpoints.

At 0.6B, the comparison between D and D-remake is less uniform. D-remake has lower mean Suite II loss than D at 5k, whereas D achieves lower mean Suite I loss and higher mean downstream scores at 11k. Thus, the relative effects of the two rewriting instructions differ across evaluation suites, training stages, and model settings.

These results demonstrate that substantial behavioral recovery is possible after an early deficit, but its extent varies across model settings and rewriting instructions. The prompt comparison does not isolate the causal effect of simplification because changing the instruction may alter multiple properties of the rewritten corpus.

Spectral comparison. We analyze weight-matrix spectra at both checkpoints using the same diagnostic procedure within each scale. Table 15 reports the proportion of analyzed matrices with $\alpha \in [2, 6]$. Each model contributes 224 analyzed matrices.

At 0.6B, A increases from 159/224 at 5k to 174/224 at 11k, while D changes from 163/224 to 173/224. Thus, the final aggregate difference between D and A is only one matrix.

At 3B, A changes from 197/224 to 195/224, while D changes from 195/224 to 191/224. D-remake changes from 192/224 to 191/224. The two rewriting variants therefore have identical final aggregate proportions, despite their substantially different held-out loss profiles.

Table 15: Aggregate spectral proportions in the 200B-token cross-scale experiments. Entries give the number and percentage of analyzed matrices with $\alpha \in [2, 6]$. Each model has 224 analyzed matrices. Spectral results for 0.6B D-remake are not reported.

Scale	Condition	5k	11k
0.6B	A	159 (70.98%)	174 (77.68%)
	D	163 (72.77%)	173 (77.23%)
3B	A	197 (87.95%)	195 (87.05%)
	D	195 (87.05%)	191 (85.27%)
	D-remake	192 (85.71%)	191 (85.27%)

The aggregate proportions differ considerably between the two model scales, and their interpretation depends on architecture and the fixed α interval. They should not be treated as scale-independent measures of model quality or generalization. Although layer-wise diagnostics provide additional structural information, baseline reruns reveal substantial variability in some localized spectral statistics. The small final aggregate differences observed here therefore do not establish reproducible intervention-specific structural effects. The corresponding relationships with behavioral recovery remain descriptive.

Consistency with the longer-budget experiment. For the 0.6B model, the intervention-phase held-out loss differences show substantial correspondence between the 200B experiment at 5k and the original 1.2T experiment at 20k. Across individual datasets, the correlations of D–A differences are approximately 0.87 on Suite I and 0.73 on Suite II.

This consistency is not complete across evaluation types. The mean downstream score difference is -0.0076 at 200B/5k, compared with +0.018 in the original 1.2T/20k experiment. The respective Suite II mean loss differences are +0.0412 and approximately +0.003. Thus, the results support partial replication of the held-out loss patterns, rather than equivalence of the full behavioral effects across token budgets.

Sensitivity of Suite II comparisons. Two English subsets, `English_reading_easy` and `high_freq_en`, exert a substantial influence on the 0.6B final Suite II mean difference. Excluding these subsets changes the 0.6B D–A mean from -0.0083 across 16 datasets to +0.0038 across the remaining 14 datasets, with 8 of 14 showing lower loss under D. At 5k, the corresponding 14-dataset mean is +0.0517, with 2 of 14 showing lower loss.

The same exclusion leaves the 3B D–A mean positive at both checkpoints (+0.2287 at 5k and +0.0565 at 11k). The broad early deficit and incomplete final recovery at 3B are therefore not explained by these two subsets alone.

Interpretation and limitations. The cross-scale experiments establish several descriptive observations: rewriting can induce different intervention-phase behavioral effects at different model settings; substantial reductions in some deficits are observed between 5k and 11k; and recovery can differ markedly between rewriting instructions. These effects are not captured uniquely by the aggregate $\alpha \in [2, 6]$ proportion.

Each condition is represented by one training run, without independent intervention-seed replication. The two checkpoints provide a coarse view of behavioral and structural evolution rather than a complete recovery trajectory. Because 5k precedes the exact phase transition, the changes observed between 5k and 11k cannot be attributed exclusively to common-data training. Moreover, although training configurations are matched within each scale, the comparison between 0.6B and 3B does not isolate parameter count from other architecture-dependent properties. No separate intervention mechanism is identified from these observations.

Detailed cross-scale behavioral results. The following tables report all individual downstream scores and held-out losses for Conditions A, D, and D-remake at 5k and 11k steps. For consistent presentation, held-out dataset names use the suffix `-dev`, and downstream metric names include their category prefixes. Abbreviations D-r and Δ_{D-r} denote D-remake and its difference from A. Differences are recalculated from the displayed model values and may differ from the source

difference columns by the last decimal place due to rounding. Positive downstream-score differences and negative held-out-loss differences favor the corresponding intervention.

Table 16: 0.6B downstream scores at 5k and 11k, including A, D, D-remake (D-r), and their pairwise differences (higher is better).

Step	Metric / dataset	A	D	D-r	$D - A$	$D\text{-}r - A$	$D\text{-}r - D$
5k	Coding/Open-Source_text_humaneval	0.1159	0.1037	0.1098	−0.0122	−0.0061	+0.0061
5k	General_Text/agieval_acc	0.0803	0.1040	0.0951	+0.0237	+0.0148	−0.0089
5k	General_Text/arc_c_acc	0.2475	0.2678	0.2610	+0.0203	+0.0135	−0.0068
5k	General_Text/arc_e_acc	0.2557	0.2416	0.2875	−0.0141	+0.0318	+0.0459
5k	General_Text/boolq_en	0.5428	0.4654	0.5627	−0.0774	+0.0199	+0.0973
5k	General_Text/cmmlu_acc	0.2705	0.2567	0.2568	−0.0138	−0.0137	+0.0001
5k	General_Text/hellaswagen_acc_norm	0.3131	0.3079	0.3061	−0.0052	−0.0070	−0.0018
5k	General_Text/math23k_cot_acc	0.0792	0.0822	0.0812	+0.0030	+0.0020	−0.0010
5k	General_Text/piqa_en_acc_norm	0.6289	0.6132	0.6143	−0.0157	−0.0146	+0.0011
5k	General_Text/winogrande_en_acc_norm	0.4901	0.5004	0.5375	+0.0103	+0.0474	+0.0371
5k	Knowledge&Reasoning/Open-Source_text_bbhen	0.2576	0.2615	0.2706	+0.0039	+0.0130	+0.0091
5k	Knowledge&Reasoning/Open-Source_text_c3	0.2849	0.3219	0.2630	+0.0370	−0.0219	−0.0589
5k	Knowledge&Reasoning/Open-Source_text_ceval	0.2400	0.2296	0.2303	−0.0104	−0.0097	+0.0007
5k	Knowledge&Reasoning/Open-Source_text_drop	0.2066	0.1725	0.1162	−0.0341	−0.0904	−0.0563
5k	Knowledge&Reasoning/Open-Source_text_gpqa_avg@4	0.2428	0.2360	0.2076	−0.0068	−0.0352	−0.0284
5k	Knowledge&Reasoning/Open-Source_text_gpqa_pass@4	0.6116	0.5804	0.4665	−0.0312	−0.1451	−0.1139
5k	Knowledge&Reasoning/Open-Source_text_gsm8ken	0.1759	0.1736	0.1433	−0.0023	−0.0326	−0.0303
5k	Knowledge&Reasoning/Open-Source_text_mathen-fewshot_avg@4	0.0903	0.0864	0.0830	−0.0039	−0.0073	−0.0034
5k	Knowledge&Reasoning/Open-Source_text_mathen-fewshot_pass@4	0.2102	0.2030	0.1992	−0.0072	−0.0110	−0.0038
5k	Knowledge&Reasoning/Open-Source_text_mmluen	0.2760	0.2607	0.2293	−0.0153	−0.0467	−0.0314
11k	Coding/Open-Source_text_humaneval	0.1738	0.1585	0.1524	−0.0153	−0.0214	−0.0061
11k	General_Text/agieval_acc	0.0849	0.0925	0.0851	+0.0076	+0.0002	−0.0074
11k	General_Text/arc_c_acc	0.2576	0.2915	0.2576	+0.0339	0.0000	−0.0339
11k	General_Text/arc_e_acc	0.2734	0.2716	0.2892	−0.0018	+0.0158	+0.0176
11k	General_Text/boolq_en	0.4740	0.4468	0.5875	−0.0272	+0.1135	+0.1407
11k	General_Text/cmmlu_acc	0.2760	0.2718	0.2829	−0.0042	+0.0069	+0.0111
11k	General_Text/hellaswagen_acc_norm	0.3386	0.3381	0.3289	−0.0005	−0.0097	−0.0092
11k	General_Text/math23k_cot_acc	0.1274	0.1334	0.1404	+0.0060	+0.0130	+0.0070
11k	General_Text/piqa_en_acc_norm	0.6366	0.6453	0.6382	+0.0087	+0.0016	−0.0071
11k	General_Text/winogrande_en_acc_norm	0.5122	0.5107	0.5146	−0.0015	+0.0024	+0.0039
11k	Knowledge&Reasoning/Open-Source_text_bbhen	0.2819	0.2845	0.2845	+0.0026	+0.0026	0.0000
11k	Knowledge&Reasoning/Open-Source_text_c3	0.3219	0.3397	0.3096	+0.0178	−0.0123	−0.0301
11k	Knowledge&Reasoning/Open-Source_text_ceval	0.2266	0.2303	0.2318	+0.0037	+0.0052	+0.0015
11k	Knowledge&Reasoning/Open-Source_text_drop	0.2263	0.2218	0.1627	−0.0045	−0.0636	−0.0591
11k	Knowledge&Reasoning/Open-Source_text_gpqa_avg@4	0.2455	0.2539	0.2333	+0.0084	−0.0122	−0.0206
11k	Knowledge&Reasoning/Open-Source_text_gpqa_pass@4	0.6317	0.6629	0.5446	+0.0312	−0.0871	−0.1183
11k	Knowledge&Reasoning/Open-Source_text_gsm8ken	0.2616	0.3017	0.2532	+0.0401	−0.0084	−0.0485
11k	Knowledge&Reasoning/Open-Source_text_mathen-fewshot_avg@4	0.1429	0.1379	0.1368	−0.0050	−0.0061	−0.0011
11k	Knowledge&Reasoning/Open-Source_text_mathen-fewshot_pass@4	0.2906	0.2840	0.2818	−0.0066	−0.0088	−0.0022
11k	Knowledge&Reasoning/Open-Source_text_mmluen	0.2870	0.2517	0.2367	−0.0353	−0.0503	−0.0150

Table 17: 0.6B held-out loss, Suite I at 5k and 11k, including A, D, D-remake (D-r), and their pairwise differences (lower is better).

Step	Metric / dataset	A	D	D-r	$D - A$	$D\text{-}r - A$	$D\text{-}r - D$
5k	pure_code-dev	1.6302	1.6019	1.6226	−0.0283	−0.0076	+0.0207
5k	pure_code_c++-dev	1.3805	1.3720	1.3812	−0.0085	+0.0007	+0.0092
5k	pure_code_java-dev	1.2317	1.2231	1.2260	−0.0086	−0.0057	+0.0029
5k	pure_code_javascript-dev	1.6046	1.5443	1.5902	−0.0603	−0.0144	+0.0459
5k	pure_code_python-dev	1.4040	1.4084	1.4081	+0.0044	+0.0041	−0.0003
5k	pure_en_math-dev	1.5720	1.5923	1.6248	+0.0203	+0.0528	+0.0325
5k	pure_en_news-dev	3.2970	3.3280	3.3241	+0.0310	+0.0271	−0.0039
5k	pure_en_novel-dev	3.6844	3.7131	3.7140	+0.0287	+0.0296	+0.0009
5k	pure_en_qa-dev	3.1934	3.2295	3.2500	+0.0361	+0.0566	+0.0205
5k	pure_en_world_knowledge-dev	2.4831	2.5122	2.5310	+0.0291	+0.0479	+0.0188
5k	pure_hungarian_finals_exam-dev	1.1315	1.1212	1.1231	−0.0103	−0.0084	+0.0019
5k	pure_skywork_gsm8k-dev	0.9765	0.9787	0.9815	+0.0022	+0.0050	+0.0028
5k	pure_zh_math-dev	1.8374	1.8485	1.8328	+0.0111	−0.0046	−0.0157
5k	pure_zh_news-dev	3.6275	3.6566	3.6202	+0.0291	−0.0073	−0.0364
5k	pure_zh_novel-dev	4.0238	4.0186	4.0128	−0.0052	−0.0110	−0.0058
5k	pure_zh_qa-dev	4.2539	4.2770	4.2737	+0.0231	+0.0198	−0.0033
5k	pure_zh_world_knowledge-dev	3.5755	3.6313	3.5938	+0.0558	+0.0183	−0.0375
11k	pure_code-dev	1.51982	1.50626	1.51162	−0.01356	−0.00820	+0.00536
11k	pure_code_c++-dev	1.28706	1.27678	1.28711	−0.01028	+0.00005	+0.01033
11k	pure_code_java-dev	1.12429	1.11777	1.11344	−0.00652	−0.01085	−0.00433
11k	pure_code_javascript-dev	1.47669	1.45357	1.46285	−0.02312	−0.01384	+0.00928
11k	pure_code_python-dev	1.30502	1.29994	1.30318	−0.00508	−0.00184	+0.00324
11k	pure_en_math-dev	1.47748	1.47222	1.51542	−0.00526	+0.03794	+0.04320
11k	pure_en_news-dev	3.13715	3.14054	3.15924	+0.00339	+0.02209	+0.01870
11k	pure_en_novel-dev	3.54969	3.55032	3.56986	+0.00063	+0.02017	+0.01954
11k	pure_en_qa-dev	3.07802	3.07613	3.12774	−0.00189	+0.04972	+0.05161
11k	pure_en_world_knowledge-dev	2.35297	2.34544	2.37936	−0.00753	+0.02639	+0.03392
11k	pure_hungarian_finals_exam-dev	1.00839	1.00174	1.04406	−0.00665	+0.03567	+0.04232
11k	pure_skywork_gsm8k-dev	0.93175	0.95035	0.94522	+0.01860	+0.01347	−0.00513
11k	pure_zh_math-dev	1.70376	1.69865	1.70971	−0.00511	+0.00595	+0.01106
11k	pure_zh_news-dev	3.44674	3.43757	3.43901	−0.00917	−0.00773	+0.00144
11k	pure_zh_novel-dev	3.84723	3.83061	3.83179	−0.01662	−0.01544	+0.00118
11k	pure_zh_qa-dev	4.08588	4.06927	4.10709	−0.01661	+0.02121	+0.03782
11k	pure_zh_world_knowledge-dev	3.39671	3.39073	3.42019	−0.00598	+0.02348	+0.02946

Table 18: 0.6B held-out loss, Suite II at 5k and 11k, including A, D, D-remake (D-r), and their pairwise differences (lower is better).

Step	Metric / dataset	A	D	D-r	$D - A$	$D\text{-}r - A$	$D\text{-}r - D$
5k	Acient_Chinese_reading_2025_GK-dev	5.1310	5.1840	5.0978	+0.0530	−0.0332	−0.0862
5k	English_reading_easy-dev	3.1431	3.1041	3.1437	−0.0390	+0.0006	+0.0396
5k	GRE_202501-dev	3.7937	3.9143	3.8704	+0.1206	+0.0767	−0.0439
5k	HSK5-dev	4.0138	4.0149	4.0256	+0.0011	+0.0118	+0.0107
5k	Morden_Chinese_reading_2025_GK-dev	4.5607	4.6222	4.5559	+0.0615	−0.0048	−0.0663
5k	high_freq_cn-dev	4.2355	4.2871	4.2166	+0.0516	−0.0189	−0.0705
5k	high_freq_cn_sampled-dev	4.2906	4.3189	4.2638	+0.0283	−0.0268	−0.0551
5k	high_freq_en-dev	3.1856	3.1602	3.1869	−0.0254	+0.0013	+0.0267
5k	history_geography_politics_2025_GK-dev	4.0574	4.0923	3.9976	+0.0349	−0.0598	−0.0947
5k	low_freq_cn-dev	4.2900	4.2871	4.1844	−0.0029	−0.1056	−0.1027
5k	low_freq_en-dev	3.5928	3.7000	3.6181	+0.1072	+0.0253	−0.0819
5k	low_freq_en_sampled-dev	3.7979	3.9155	3.8508	+0.1176	+0.0529	−0.0647
5k	news_cn-dev	3.7253	3.7442	3.6547	+0.0189	−0.0706	−0.0895
5k	news_en-dev	3.4792	3.5291	3.4903	+0.0499	+0.0111	−0.0388
5k	professional_cn-dev	3.6421	3.5990	3.4631	−0.0431	−0.1790	−0.1359
5k	professional_en-dev	2.8331	2.9589	2.7731	+0.1258	−0.0600	−0.1858
11k	Acient_Chinese_reading_2025_GK-dev	4.86521	4.88166	4.88336	+0.01645	+0.01815	+0.00170
11k	English_reading_easy-dev	3.11377	3.01526	3.04230	−0.09851	−0.07147	+0.02704
11k	GRE_202501-dev	3.61122	3.66301	3.72367	+0.05179	+0.11245	+0.06066
11k	HSK5-dev	3.78652	3.79553	3.83077	+0.00901	+0.04425	+0.03524
11k	Morden_Chinese_reading_2025_GK-dev	4.33092	4.32687	4.33231	−0.00405	+0.00139	+0.00544
11k	high_freq_cn-dev	4.02727	4.01540	4.03098	−0.01187	+0.00371	+0.01558
11k	high_freq_cn_sampled-dev	4.08013	4.06654	4.09382	−0.01359	+0.01369	+0.02728

Continued on next page

Continued from previous page

Step	Metric / dataset	A	D	D-r	$D - A$	$D\text{-}r - A$	$D\text{-}r - D$
11k	high_freq_en-dev	3.15054	3.06220	3.08891	−0.08834	−0.06163	+0.02671
11k	history_geography_politics_2025_GK-dev	3.85775	3.85650	3.82704	−0.00125	−0.03071	−0.02946
11k	low_freq_cn-dev	3.99961	3.99187	3.93360	−0.00774	−0.06601	−0.05827
11k	low_freq_en-dev	3.39296	3.42558	3.44536	+0.03262	+0.05240	+0.01978
11k	low_freq_en_sampled-dev	3.60743	3.62855	3.68328	+0.02112	+0.07585	+0.05473
11k	news_cn-dev	3.51528	3.49865	3.49713	−0.01663	−0.01815	−0.00152
11k	news_en-dev	3.35619	3.35214	3.36164	−0.00405	+0.00545	+0.00950
11k	professional_cn-dev	3.35385	3.32492	3.17822	−0.02893	−0.17563	−0.14670
11k	professional_en-dev	2.60655	2.61739	2.54325	+0.01084	−0.06330	−0.07414

Table 19: 3B downstream scores at 5k and 11k, including A, D, D-remake (D-r), and their pairwise differences (higher is better).

Step	Metric / dataset	A	D	D-r	$D - A$	$D\text{-}r - A$	$D\text{-}r - D$
5k	Coding/Open-Source_text_humaneval	0.1939	0.1402	0.1098	−0.0537	−0.0841	−0.0304
5k	General_Text/agieval_acc	0.1126	0.0977	0.0765	−0.0149	−0.0361	−0.0212
5k	General_Text/arc_c_acc	0.3153	0.2983	0.2780	−0.0170	−0.0373	−0.0203
5k	General_Text/arc_e_acc	0.3792	0.2787	0.2222	−0.1005	−0.1570	−0.0565
5k	General_Text/boolq_en	0.6107	0.6171	0.5223	+0.0064	−0.0884	−0.0948
5k	General_Text/cmmlu_acc	0.3319	0.2897	0.2956	−0.0422	−0.0363	+0.0059
5k	General_Text/hellaswagen_acc_norm	0.3959	0.3683	0.3642	−0.0276	−0.0317	−0.0041
5k	General_Text/math23k_cot_acc	0.2166	0.1514	0.1675	−0.0652	−0.0491	+0.0161
5k	General_Text/piqa_en_acc_norm	0.6670	0.6485	0.6415	−0.0185	−0.0255	−0.0070
5k	General_Text/winogrande_en_acc_norm	0.5217	0.5146	0.5541	−0.0071	+0.0324	+0.0395
5k	Knowledge&Reasoning/Open-Source_text_bbhen	0.3819	0.3241	0.3083	−0.0578	−0.0736	−0.0158
5k	Knowledge&Reasoning/Open-Source_text_c3	0.3370	0.3808	0.3068	+0.0438	−0.0302	−0.0740
5k	Knowledge&Reasoning/Open-Source_text_ceval	0.2868	0.2340	0.2296	−0.0528	−0.0572	−0.0044
5k	Knowledge&Reasoning/Open-Source_text_drop	0.2952	0.2728	0.2626	−0.0224	−0.0326	−0.0102
5k	Knowledge&Reasoning/Open-Source_text_gpqa_avg@4	0.2539	0.2466	0.2160	−0.0073	−0.0379	−0.0306
5k	Knowledge&Reasoning/Open-Source_text_gpqa_pass@4	0.6942	0.6786	0.5513	−0.0156	−0.1429	−0.1273
5k	Knowledge&Reasoning/Open-Source_text_gsm8ken	0.4761	0.3442	0.3495	−0.1319	−0.1266	+0.0053
5k	Knowledge&Reasoning/Open-Source_text_mathen-fewshot_avg@4	0.2276	0.1686	0.1542	−0.0590	−0.0734	−0.0144
5k	Knowledge&Reasoning/Open-Source_text_mathen-fewshot_pass@4	0.3946	0.3176	0.2986	−0.0770	−0.0960	−0.0190
5k	Knowledge&Reasoning/Open-Source_text_mmluen	0.3047	0.3013	0.3013	−0.0034	−0.0034	0.0000
11k	Coding/Open-Source_text_humaneval	0.2073	0.1768	0.2256	−0.0305	+0.0183	+0.0488
11k	General_Text/agieval_acc	0.1110	0.0891	0.1080	−0.0219	−0.0030	+0.0189
11k	General_Text/arc_c_acc	0.3492	0.3797	0.2848	+0.0305	−0.0644	−0.0949
11k	General_Text/arc_e_acc	0.4991	0.4392	0.4515	−0.0599	−0.0476	+0.0123
11k	General_Text/boolq_en	0.6410	0.6141	0.6025	−0.0269	−0.0385	−0.0116
11k	General_Text/cmmlu_acc	0.4158	0.3722	0.3475	−0.0436	−0.0683	−0.0247
11k	General_Text/hellaswagen_acc_norm	0.4445	0.4336	0.4192	−0.0109	−0.0253	−0.0144
11k	General_Text/math23k_cot_acc	0.3220	0.2989	0.2588	−0.0231	−0.0632	−0.0401
11k	General_Text/piqa_en_acc_norm	0.6795	0.6861	0.6687	+0.0066	−0.0108	−0.0174
11k	General_Text/winogrande_en_acc_norm	0.5375	0.5185	0.5580	−0.0190	+0.0205	+0.0395
11k	Knowledge&Reasoning/Open-Source_text_bbhen	0.4680	0.4635	0.4036	−0.0045	−0.0644	−0.0599
11k	Knowledge&Reasoning/Open-Source_text_c3	0.4192	0.4480	0.4123	+0.0288	−0.0069	−0.0357
11k	Knowledge&Reasoning/Open-Source_text_ceval	0.3395	0.2734	0.2600	−0.0661	−0.0795	−0.0134
11k	Knowledge&Reasoning/Open-Source_text_drop	0.3377	0.3115	0.2383	−0.0262	−0.0994	−0.0732
11k	Knowledge&Reasoning/Open-Source_text_gpqa_avg@4	0.2612	0.2494	0.2450	−0.0118	−0.0162	−0.0044
11k	Knowledge&Reasoning/Open-Source_text_gpqa_pass@4	0.7054	0.6161	0.6496	−0.0893	−0.0558	+0.0335
11k	Knowledge&Reasoning/Open-Source_text_gsm8ken	0.6058	0.5770	0.5178	−0.0288	−0.0880	−0.0592
11k	Knowledge&Reasoning/Open-Source_text_mathen-fewshot_avg@4	0.3118	0.3009	0.2606	−0.0109	−0.0512	−0.0403

Continued on next page

Continued from previous page

Step	Metric / dataset	A	D	D-r	$D - A$	$D\text{-}r - A$	$D\text{-}r - D$
11k	`Knowledge&Reasoning/Open-Source_text_mathen-fewshot_pass@4`	0.4886	0.4792	0.4422	−0.0094	−0.0464	−0.0370
11k	`Knowledge&Reasoning/Open-Source_text_mmluen`	0.3873	0.3550	0.3350	−0.0323	−0.0523	−0.0200

Table 20: 3B held-out loss, Suite I at 5k and 11k, including A, D, D-remake (D-r), and their pairwise differences (lower is better).

Step	Metric / dataset	A	D	D-r	$D - A$	$D\text{-}r - A$	$D\text{-}r - D$
5k	`pure_code-dev`	1.3815	1.4568	1.4657	+0.0753	+0.0842	+0.0089
5k	`pure_code_c++-dev`	1.1709	1.2469	1.2439	+0.0760	+0.0730	−0.0030
5k	`pure_code_java-dev`	1.0096	1.0860	1.0869	+0.0764	+0.0773	+0.0009
5k	`pure_code_javascript-dev`	1.3118	1.3794	1.4047	+0.0676	+0.0929	+0.0253
5k	`pure_code_python-dev`	1.1654	1.2589	1.2476	+0.0935	+0.0822	−0.0113
5k	`pure_en_math-dev`	1.3480	1.4458	1.4603	+0.0978	+0.1123	+0.0145
5k	`pure_en_news-dev`	2.9239	3.0683	3.0542	+0.1444	+0.1303	−0.0141
5k	`pure_en_novel-dev`	3.3453	3.5126	3.5036	+0.1673	+0.1583	−0.0090
5k	`pure_en_qa-dev`	2.9090	3.0549	3.0628	+0.1459	+0.1538	+0.0079
5k	`pure_en_world_knowledge-dev`	2.1632	2.2744	2.2764	+0.1112	+0.1132	+0.0020
5k	`pure_hungarian_finals_exam-dev`	0.9082	0.9656	0.9650	+0.0574	+0.0568	−0.0006
5k	`pure_skywork_gsm8k-dev`	0.9302	0.9312	0.9238	+0.0010	−0.0064	−0.0074
5k	`pure_zh_math-dev`	1.5430	1.6619	1.6518	+0.1189	+0.1088	−0.0101
5k	`pure_zh_news-dev`	3.2026	3.3921	3.3470	+0.1895	+0.1444	−0.0451
5k	`pure_zh_novel-dev`	3.6119	3.7586	3.7407	+0.1467	+0.1288	−0.0179
5k	`pure_zh_qa-dev`	3.8518	4.0357	4.0247	+0.1839	+0.1729	−0.0110
5k	`pure_zh_world_knowledge-dev`	3.1672	3.3863	3.3452	+0.2191	+0.1780	−0.0411
11k	`pure_code-dev`	1.2767	1.3016	1.3468	+0.0249	+0.0701	+0.0452
11k	`pure_code_c++-dev`	1.0923	1.1133	1.1508	+0.0210	+0.0585	+0.0375
11k	`pure_code_java-dev`	0.9224	0.9414	0.9773	+0.0190	+0.0549	+0.0359
11k	`pure_code_javascript-dev`	1.1941	1.2218	1.2738	+0.0277	+0.0797	+0.0520
11k	`pure_code_python-dev`	1.0449	1.0881	1.1290	+0.0432	+0.0841	+0.0409
11k	`pure_en_math-dev`	1.2625	1.2768	1.3582	+0.0143	+0.0957	+0.0814
11k	`pure_en_news-dev`	2.7772	2.8154	2.8717	+0.0382	+0.0945	+0.0563
11k	`pure_en_novel-dev`	3.1988	3.2475	3.3314	+0.0487	+0.1326	+0.0839
11k	`pure_en_qa-dev`	2.7931	2.8236	2.9238	+0.0305	+0.1307	+0.1002
11k	`pure_en_world_knowledge-dev`	2.0296	2.0632	2.1224	+0.0336	+0.0928	+0.0592
11k	`pure_hungarian_finals_exam-dev`	0.8905	0.8431	0.8734	−0.0474	−0.0171	+0.0303
11k	`pure_skywork_gsm8k-dev`	0.8706	0.8579	0.8931	−0.0127	+0.0225	+0.0352
11k	`pure_zh_math-dev`	1.4277	1.4545	1.5101	+0.0268	+0.0824	+0.0556
11k	`pure_zh_news-dev`	3.0302	3.0791	3.1493	+0.0489	+0.1191	+0.0702
11k	`pure_zh_novel-dev`	3.4377	3.4748	3.5439	+0.0371	+0.1062	+0.0691
11k	`pure_zh_qa-dev`	3.6894	3.7308	3.8351	+0.0414	+0.1457	+0.1043
11k	`pure_zh_world_knowledge-dev`	2.9956	3.0390	3.1447	+0.0434	+0.1491	+0.1057

Table 21: 3B held-out loss, Suite II at 5k and 11k, including A, D, D-remake (D-r), and their pairwise differences (lower is better).

Step	Metric / dataset	A	D	D-r	$D - A$	$D\text{-}r - A$	$D\text{-}r - D$
5k	`Acient_Chinese_reading_2025_GK-dev`	4.5578	4.8882	4.8130	+0.3304	+0.2552	−0.0752
5k	`English_reading_easy-dev`	2.8902	2.9323	2.9676	+0.0421	+0.0774	+0.0353
5k	`GRE_202501-dev`	3.3635	3.6237	3.5719	+0.2602	+0.2084	−0.0518
5k	`HSK5-dev`	3.5799	3.7355	3.7570	+0.1556	+0.1771	+0.0215
5k	`Morden_Chinese_reading_2025_GK-dev`	4.0716	4.3249	4.2932	+0.2533	+0.2216	−0.0317
5k	`high_freq_cn-dev`	3.7788	4.0178	3.9639	+0.2390	+0.1851	−0.0539
5k	`high_freq_cn_sampled-dev`	3.8462	4.0570	4.0413	+0.2108	+0.1951	−0.0157
5k	`high_freq_en-dev`	2.9393	2.9860	3.0171	+0.0467	+0.0778	+0.0311
5k	`history_geography_politics_2025_GK-dev`	3.5721	3.8117	3.7528	+0.2396	+0.1807	−0.0589
5k	`low_freq_cn-dev`	3.6760	3.9223	3.8628	+0.2463	+0.1868	−0.0595
5k	`low_freq_en-dev`	3.1566	3.3904	3.3358	+0.2338	+0.1792	−0.0546
5k	`low_freq_en_sampled-dev`	3.3647	3.6117	3.5610	+0.2470	+0.1963	−0.0507
5k	`news_cn-dev`	3.2778	3.4785	3.4090	+0.2007	+0.1312	−0.0695

Continued on next page

Continued from previous page

Step	Metric / dataset	A	D	D-r	$D - A$	$D\text{-}r - A$	$D\text{-}r - D$
5k	news_en-dev	3.1772	3.3029	3.2957	+0.1257	+0.1185	−0.0072
5k	professional_cn-dev	2.9687	3.1413	3.0301	+0.1726	+0.0614	−0.1112
5k	professional_en-dev	2.2387	2.5249	2.4385	+0.2862	+0.1998	−0.0864
11k	Acient_Chinese_reading_2025_GK-dev	4.2920	4.4678	4.5311	+0.1758	+0.2391	+0.0633
11k	English_reading_easy-dev	2.7956	2.8062	2.9578	+0.0106	+0.1622	+0.1516
11k	GRE_202501-dev	3.1782	3.2281	3.3615	+0.0499	+0.1833	+0.1334
11k	HSK5-dev	3.3512	3.3964	3.5403	+0.0452	+0.1891	+0.1439
11k	Morden_Chinese_reading_2025_GK-dev	3.8740	3.9167	4.0365	+0.0427	+0.1625	+0.1198
11k	high_freq_cn-dev	3.5685	3.6237	3.7309	+0.0552	+0.1624	+0.1072
11k	high_freq_cn_sampled-dev	3.6566	3.7006	3.8207	+0.0440	+0.1641	+0.1201
11k	high_freq_en-dev	2.8339	2.8449	2.9936	+0.0110	+0.1597	+0.1487
11k	history_geography_politics_2025_GK-dev	3.3629	3.4530	3.5379	+0.0901	+0.1750	+0.0849
11k	low_freq_cn-dev	3.3916	3.4922	3.5342	+0.1006	+0.1426	+0.0420
11k	low_freq_en-dev	2.9553	2.9996	3.1292	+0.0443	+0.1739	+0.1296
11k	low_freq_en_sampled-dev	3.1523	3.1879	3.3402	+0.0356	+0.1879	+0.1523
11k	news_cn-dev	3.1222	3.1542	3.2383	+0.0320	+0.1161	+0.0841
11k	news_en-dev	3.0363	3.0285	3.1503	−0.0078	+0.1140	+0.1218
11k	professional_cn-dev	2.6498	2.6629	2.6857	+0.0131	+0.0359	+0.0228
11k	professional_en-dev	1.9641	2.0340	2.1277	+0.0699	+0.1636	+0.0937

A.8 Baseline Reproducibility and Run-to-Run Variation

Behavioral variation across baseline runs. We examine baseline training variability using three independently executed Condition A runs at 20k steps. The original baseline (A) and a rerun (A-new) share the same initialization seed, whereas a third run (A-reseed) uses a different seed. All three runs are evaluated on the same 33 held-out loss datasets (17 in Suite I and 16 in Suite II) and 20 downstream metrics. Table 22 reports the unweighted mean absolute per-dataset or per-metric differences for each pair of runs.

The same-seed rerun exhibits nonzero behavioral differences from the original baseline. Comparisons involving different seeds generally show larger held-out loss differences, particularly on Suite II. However, these comparisons do not isolate the effects of initialization seeds from other sources of run-to-run variation.

At 20k, Condition C has higher held-out losses than the original A on all 33 datasets, with an unweighted mean difference of approximately +0.271. Its downstream scores are lower on 18 of 20 metrics, with a mean difference of -0.040. These early deficits are substantially larger than the observed baseline run-to-run differences, especially in held-out loss.

Condition D shows more heterogeneous behavior. Its mean signed held-out loss difference relative to A is approximately +0.004 across the 33 datasets, while its mean downstream score difference is +0.018, with improvements on 15 of 20 metrics. The small signed loss mean reflects offsetting dataset-level differences and does not imply uniformly small effects.

Spectral variation across baseline runs. At 20k, the three Condition A runs have proportions of matrices with $\alpha \in [2, 6]$ of 81.70% (183/224) for the original A, 79.46% (178/224) for A-new, and 81.25% (182/224) for A-reseed. The maximum observed difference is five out of 224 matrices, or approximately 2.2 percentage points. The same-seed rerun is not closer to the original baseline than the different-seed run in this aggregate statistic.

Layer-wise comparisons between A-new and A-reseed reveal diagnostic-dependent variation. The `mlp.up_proj` α profiles and adjacent-layer change-variance profiles exhibit broadly similar depth-wise patterns. In contrast, `self_attn.v_proj` shows substantial differences in individual spectral peaks and fitted-tail eigenvalue counts. Adjacent-layer effective-rank differences also vary considerably at selected layer positions.

These observations indicate that aggregate and layer-wise spectral statistics differ in their run-to-run sensitivity. In particular, small aggregate gaps and isolated layer-wise deviations cannot be established as reproducible intervention-specific signatures based on the available baseline comparisons.

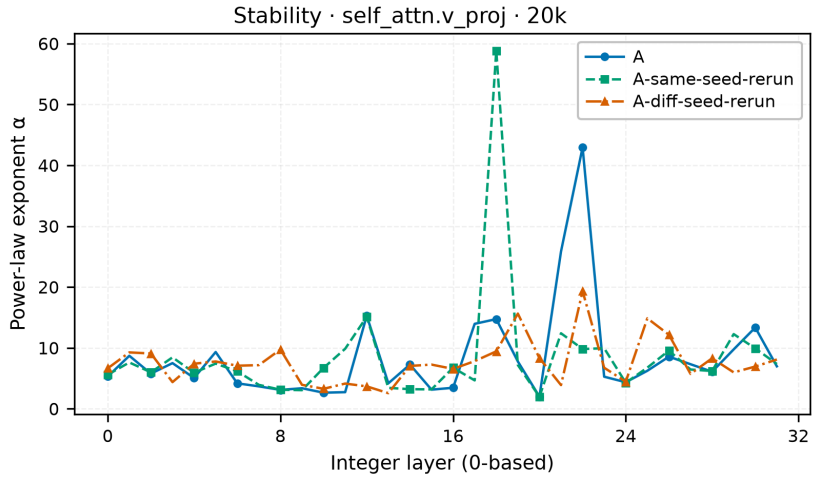
Table 22: Baseline run-to-run variation at 20k steps. Entries are mean absolute differences across 17 Suite I datasets, 16 Suite II datasets, all 33 held-out datasets, and 20 downstream metrics. The three runs represent two distinct initialization seeds.

Comparison	Suite I	Suite II	All loss	Downstream
A vs. A-new (same seed)	0.0073	0.0175	0.0123	0.0134
A vs. A-reseed (different seed)	0.0067	0.0343	0.0201	0.0162
A-new vs. A-reseed (different seed)	0.0067	0.0349	0.0204	0.0153

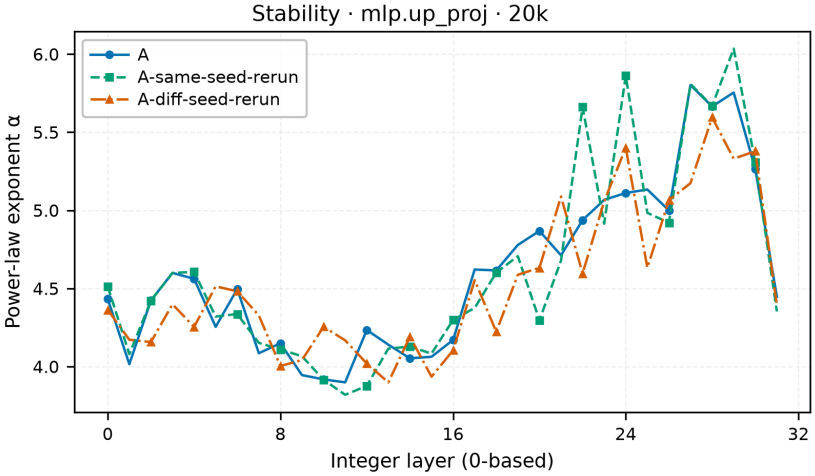
Scope and limitations. The three baseline runs provide a preliminary reference for behavioral and spectral variability at 20k. The observed differences are not statistical confidence intervals or general variability thresholds, and cannot be directly extrapolated to other checkpoints or model scales.

Available layer-wise comparisons include the original A and both retrained baselines at 20k. Conditions C and D have not been independently replicated across seeds, and their intervention-specific behavioral and parameter-space effects therefore remain unverified for cross-run reproducibility.

Spectral fitting provenance. The original 1.2T-token analysis was reported to use WeightWatcher version 0.7.6 with default fitting settings, whereas the additional baseline and 200B-token cross-scale analysis records identify version 0.7.7, also using the default fitting path. The built-in power-law fitting procedure selects the spectral-tail cutoff $x_{\min}$ by minimizing the Kolmogorov–Smirnov distance.

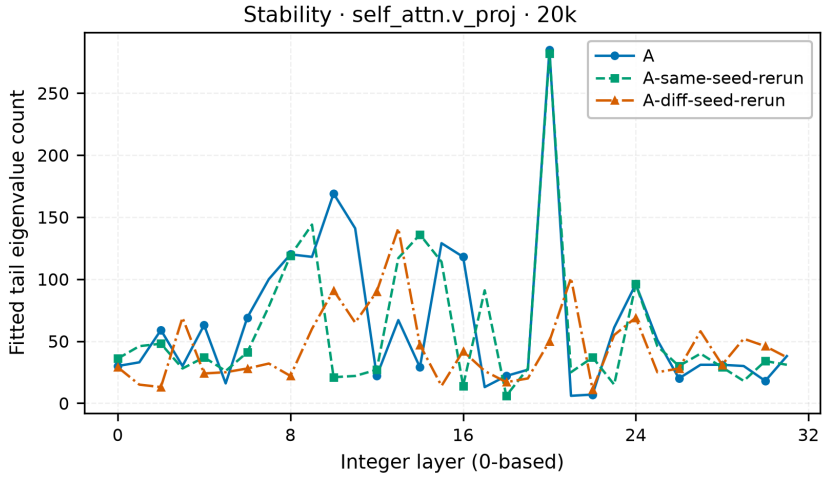


(a) Attention value projection

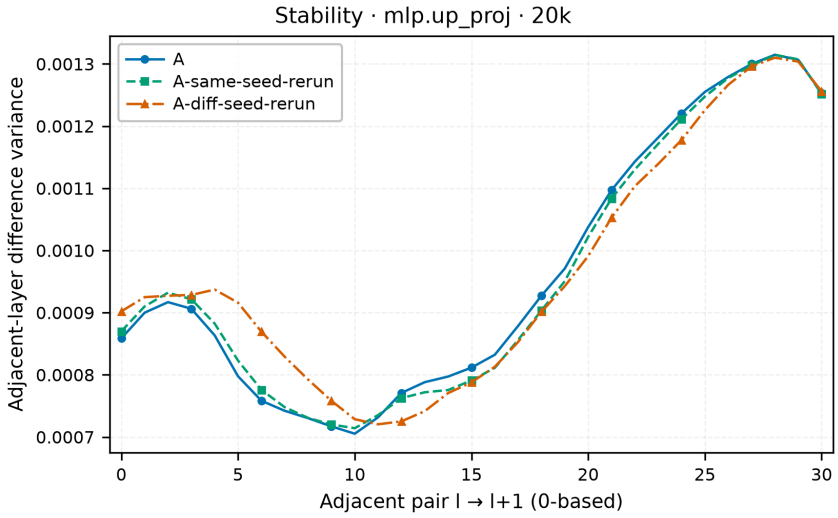


(b) MLP up projection

Figure 9: Layer-wise power-law exponents across three Condition A runs at 20k steps in the 0.6B/1.2T training setting: the original baseline (A), a same-initialization-seed rerun (A-new), and a different-initialization-seed rerun (A-reseed). Panels show (a) `self_attn.v_proj` and (b) `mlp.up_proj`. Localized attention spectral peaks vary across the observed runs, whereas the MLP exponent profiles exhibit broadly similar depth-wise patterns. These comparisons provide a descriptive reference at this checkpoint.

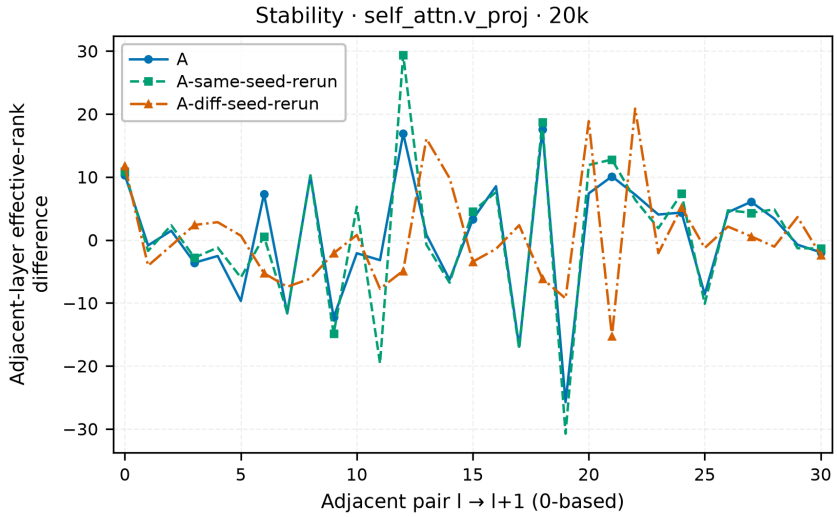


(a) Attention fitted-tail eigenvalue counts

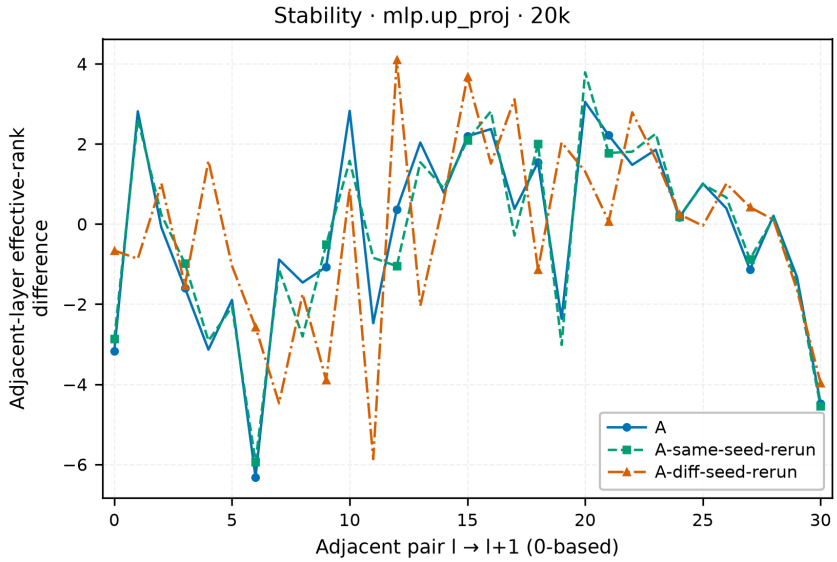


(b) MLP adjacent-layer change variance

Figure 10: Additional baseline diagnostics at 20k steps in the 0.6B/1.2T training setting. Curves show the original A, the same-seed rerun A-new, and the different-seed rerun A-reseed. Panel (a) reports the number of eigenvalues in the fitted spectral tail of `self_attn.v_proj`; these counts vary substantially at some layers. Panel (b) reports the entry-wise population variance of adjacent-layer weight differences in `mlp.up_proj`, whose raw profiles have broadly similar depth-wise shapes. Change variance is computed within each checkpoint and does not measure temporal parameter updates.



(a) Attention value projection



(b) MLP up projection

Figure 11: Adjacent-layer effective-rank differences across three Condition A runs at 20k steps in the 0.6B/1.2T training setting. Curves show the original A, the same-seed rerun A-new, and the different-seed rerun A-reseed. Panels show (a) `self_attn.v_proj` and (b) `mlp.up_proj`. Each value is computed as $r_{\text{eff}}(W_{l+1}) - r_{\text{eff}}(W_l)$ within the same checkpoint. Localized differences vary across the observed runs, providing a checkpoint-specific reference for interpreting inter-layer diagnostics.

A.9 External Validation: Additional Figures

Model-level spectral summaries. The following histograms report fitted power-law exponents for all analyzed weight-matrix records in each instruct model. Shaded regions mark $\alpha < 2$ and $\alpha > 6$. The plotted "overfit" and "underfit" labels are conventional library terminology, not diagnoses of generalization performance.

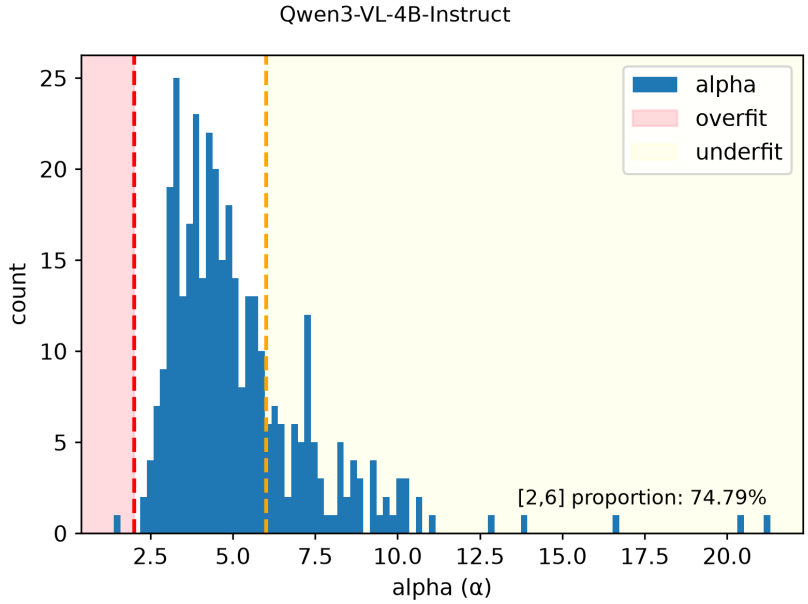


Figure 12: Model-level distribution of fitted power-law exponents for Qwen3-VL-4B-Instruct. The histogram includes 357 analyzed weight-matrix records, of which 267 (74.79%) have $\alpha \in [2, 6]$. The denominator includes the analyzed decoder and non-decoder modules under this model's inclusion rules.

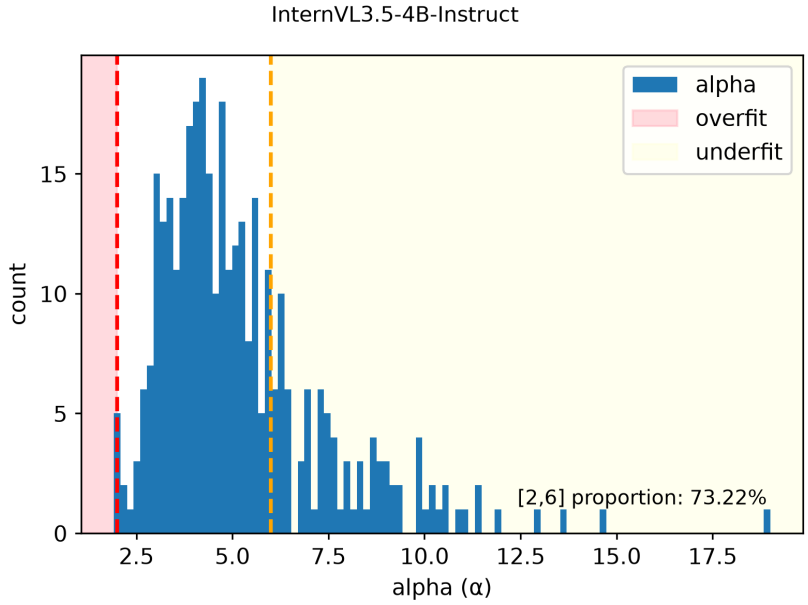


Figure 13: Model-level distribution of fitted power-law exponents for InternVL3.5-4B-Instruct. The histogram includes 351 analyzed weight-matrix records, of which 257 (73.22%) have $\alpha \in [2, 6]$. The denominator includes the analyzed decoder and non-decoder modules under this model's inclusion rules.

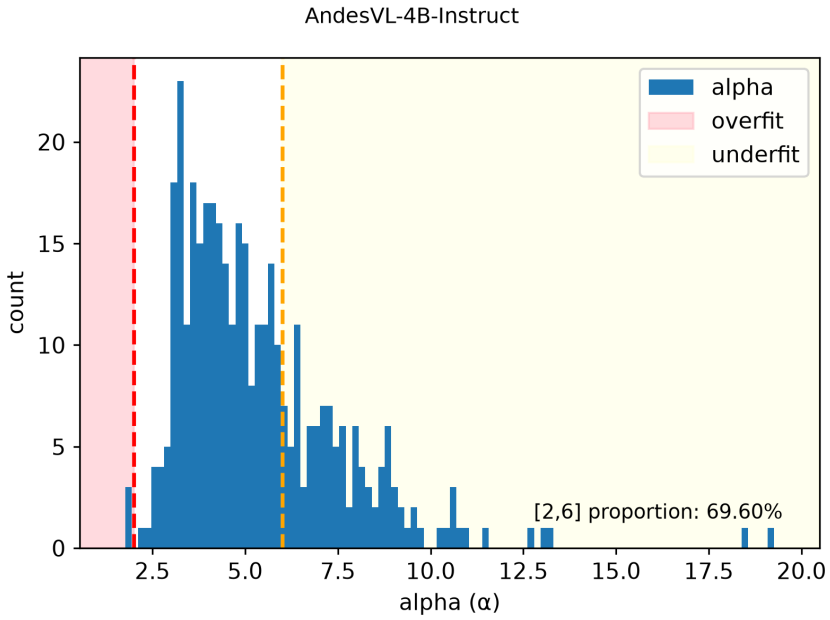
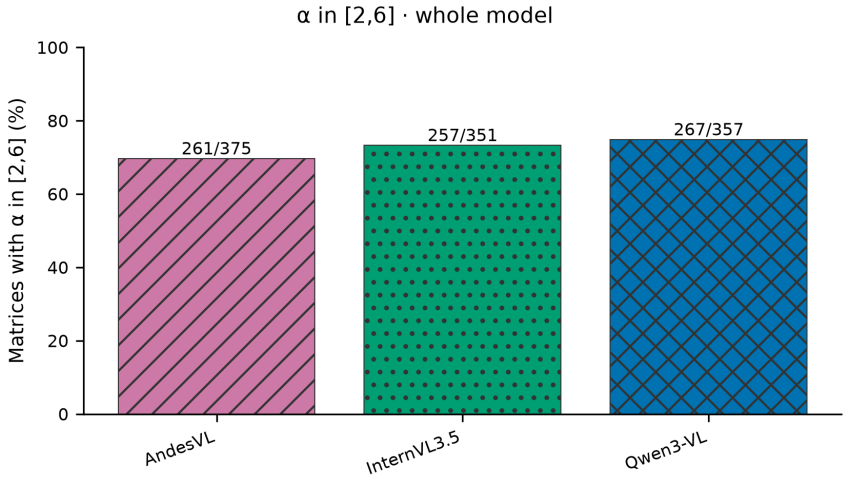
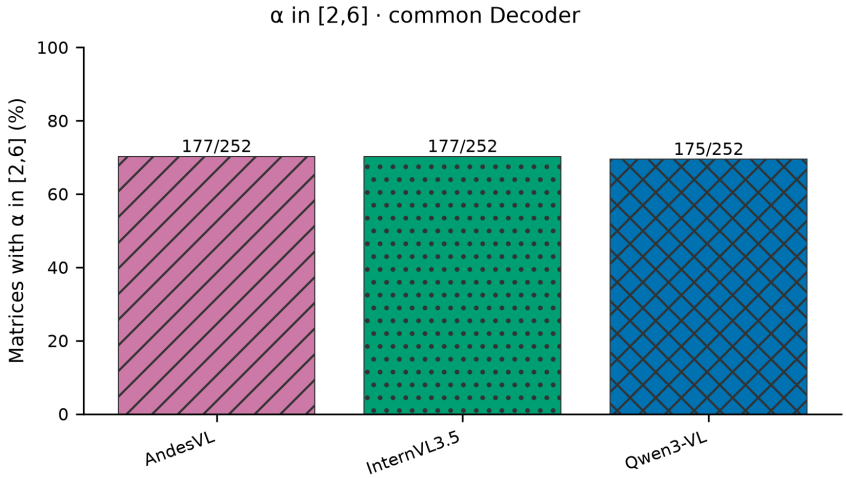


Figure 14: Model-level distribution of fitted power-law exponents for AndesVL-4B-Instruct. The histogram includes 375 analyzed weight-matrix records, of which 261 (69.60%) have $\alpha \in [2, 6]$. The denominator includes the analyzed decoder and non-decoder modules under this model's inclusion rules.



(a) Whole-model analysis inventories



(b) Common decoder matrices

Figure 15: Dependence of aggregate spectral proportions on matrix inclusion across three instruct models. Panel (a) includes all analyzed weight-matrix records under each model's module inventory. Panel (b) restricts the comparison to 252 corresponding decoder matrices per model, comprising seven projections in each of 36 layers. Bar annotations report the number of matrices with $\alpha \in [2, 6]$ divided by the number analyzed. The common-decoder proportions are similar despite differences in the whole-model summaries. These proportions are descriptive spectral statistics, not standalone model-quality scores.

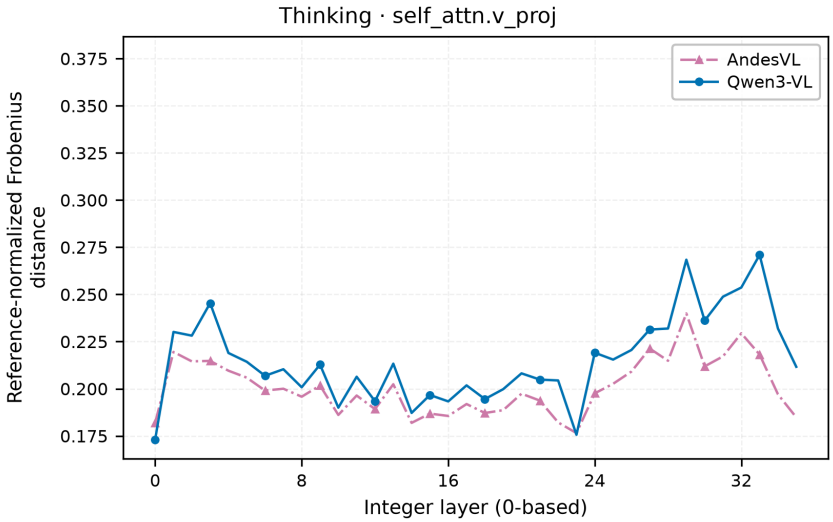


Figure 16: Relative parameter change in `self_attn.v_proj` relative to the Qwen3-4B-Base for thinking variants. Qwen3-VL-4B-Thinking (21.5%) and AndesVL-4B-Thinking (20.1%) show the same graded ordering as the instruct variants.

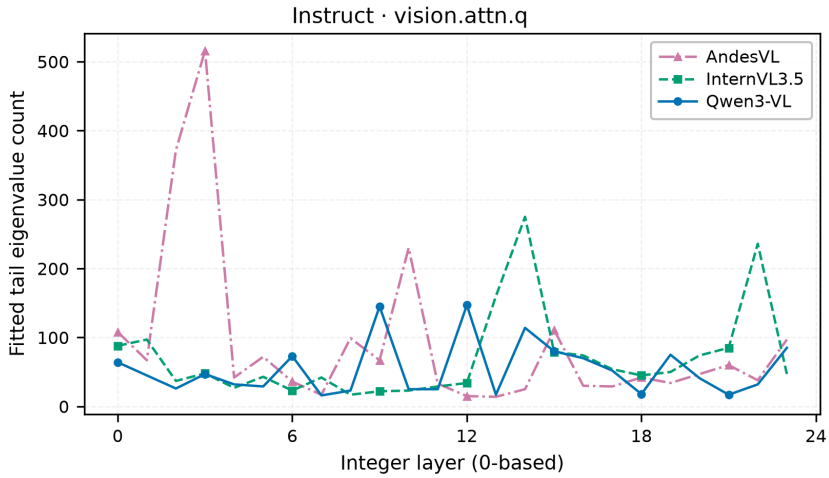
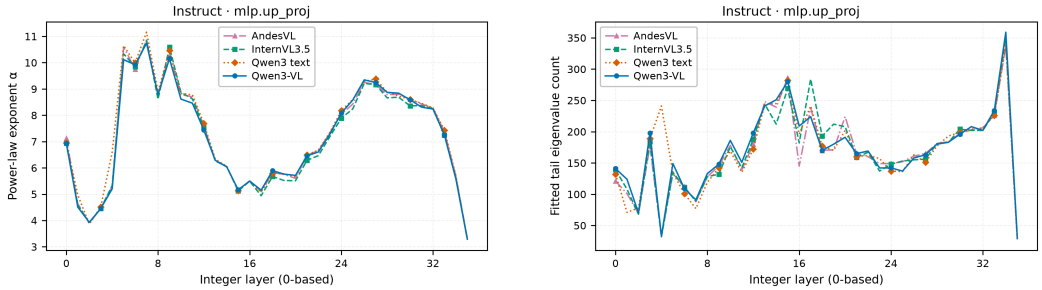


Figure 17: Layer-wise fitted-tail eigenvalue count of the vision encoder attention query weight (`attn.qkv_q`) across MLLM instruct models. AndesVL exhibits a sharp early-layer spike, supporting the interpretation that localized structural irregularity can appear in modality-specific submodules.



(a) MLP α (b) MLP fitted-tail eigenvalue counts

Figure 18: Additional MLP-pathway comparisons across instruct models. The MLP α profiles and fitted-tail eigenvalue counts exhibit greater similarity across the analyzed models than their attention-projection counterparts, although localized differences remain. This comparison supports the relative attention–MLP contrast discussed in Section 5.

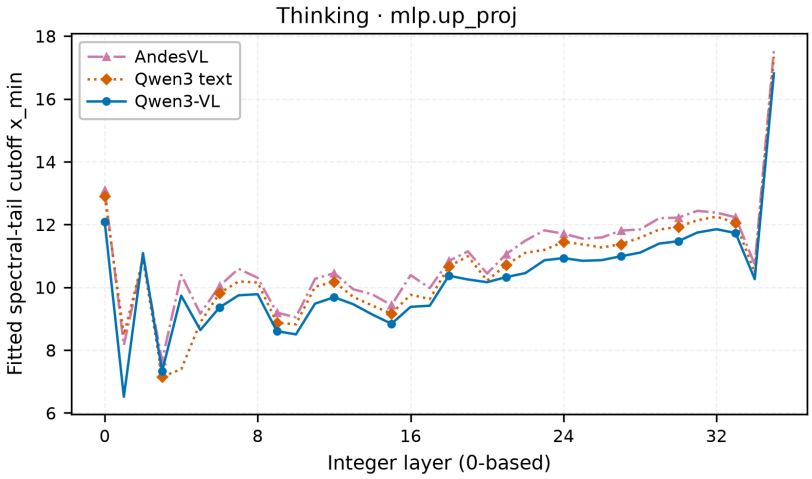


Figure 19: Layer-wise fitted spectral-tail cutoff $x_{\min}$ in `mlp.up_proj` across thinking variants. The cross-model ordering is broadly consistent with that observed in the corresponding instruct-model comparisons across most layer positions.

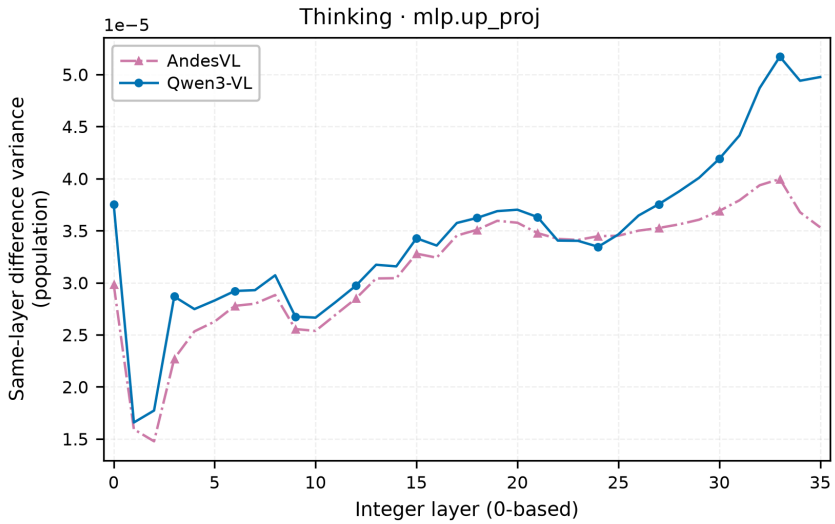


Figure 20: Same-layer weight-difference population variance in `mlp.up_proj` for the analyzed thinking variants, relative to Qwen3-4B-Base. Matrices are matched by integer layer index. This cross-model statistic differs from the within-checkpoint adjacent-layer change variance defined in Section 3.

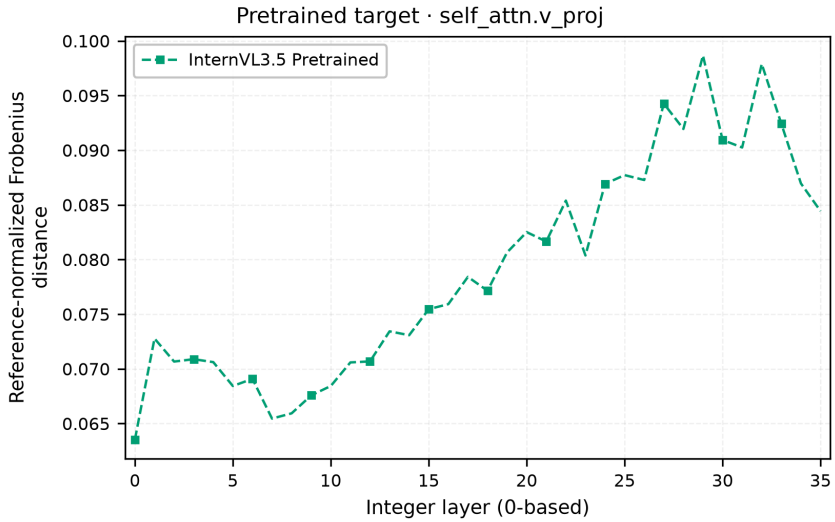
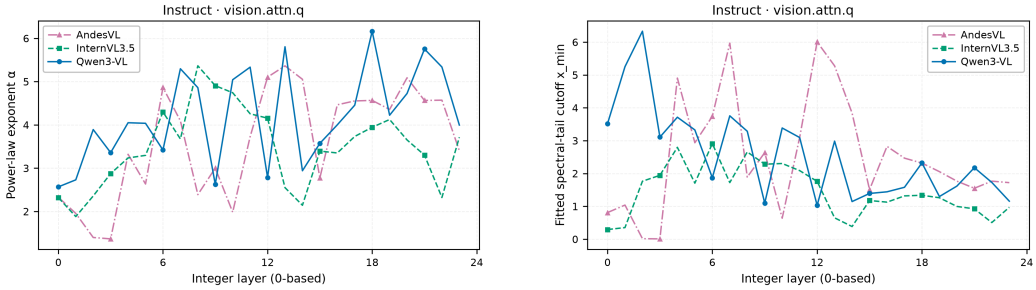


Figure 21: Same-layer reference-normalized Frobenius distance in `self_attn.v_proj` between InternVL3.5-4B-Pretrained and Qwen3-4B-Base. Matrices are matched by integer layer index, and distances are normalized by the reference matrix Frobenius norm. These comparisons do not establish updates from a verified shared initialization.



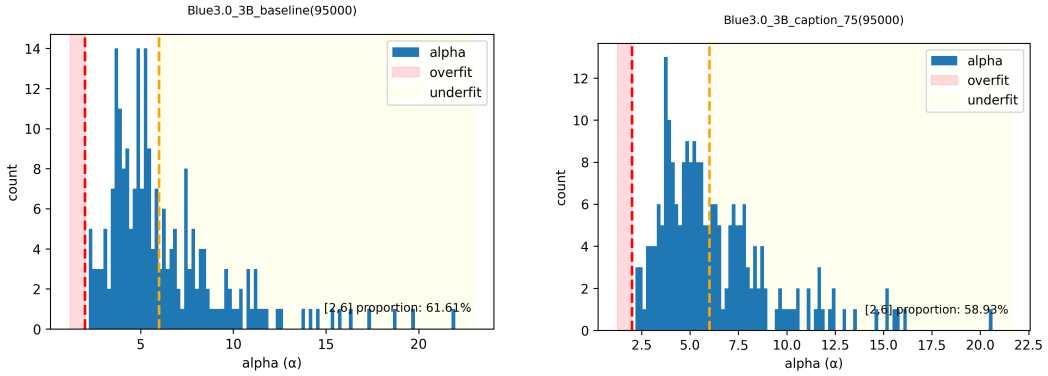
(a) Vision encoder α

(b) Vision encoder $x_{\min}$

Figure 22: Additional vision encoder diagnostics for the attention query projection. AndesVL exhibits a sharp early-layer anomaly with unusually low α and small fitted $x_{\min}$ values, supporting the interpretation in Section 5 that localized structural irregularity can appear in modality-specific submodules even when decoder modification remains comparatively conservative.

A.10 Prompt De-duplication: Additional Figures

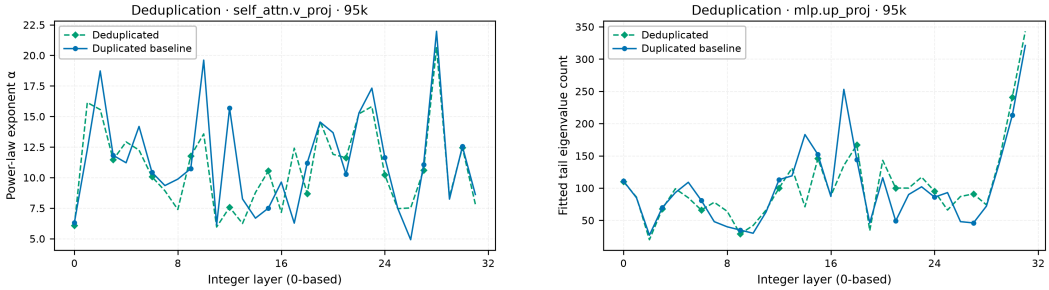
The plotted "overfit" and "underfit" labels denote conventional spectral ranges and should not be interpreted as diagnoses of generalization performance.



(a) Duplicated baseline. $\alpha \in [2, 6]$: 61.61%.

(b) De-duplicated variant. $\alpha \in [2, 6]$: 58.93%.

Figure 23: Model-level α distributions for the duplicated baseline and de-duplicated variant at 95k training steps. The reported $\alpha \in [2, 6]$ proportion decreases from 61.61% to 58.93%, while the mean scores improve in four of the five evaluation categories reported in Appendix A.3. This comparison illustrates why aggregate spectral summaries should not be interpreted as standalone quality measures.



(a) Attention α

(b) MLP fitted-tail eigenvalue count

Figure 24: Layer-wise prompt de-duplication diagnostics at 95k training steps. The duplicated baseline and de-duplicated variant exhibit localized differences in `self_attn.v_proj` α. The fitted-tail eigenvalue counts in `mlp.up_proj` show broadly similar depth-wise patterns, with noticeable differences at selected layers.

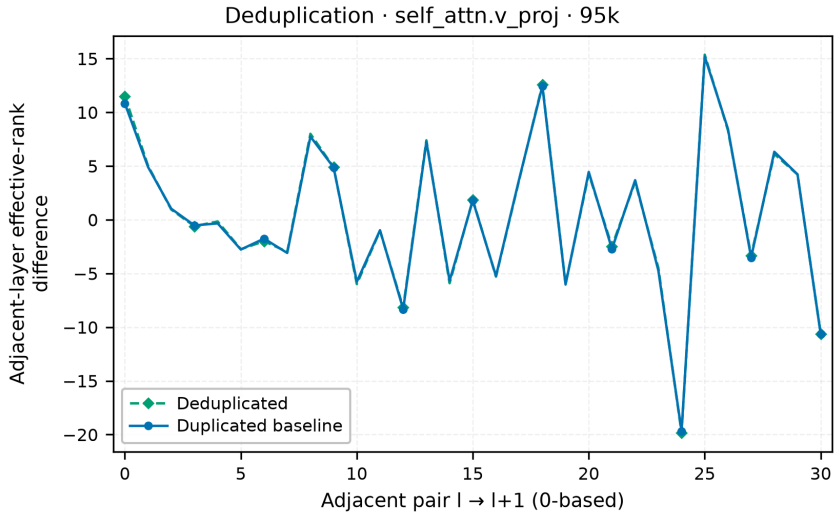


Figure 25: Adjacent-layer effective-rank differences in `self_attn.v_proj` at 95k training steps. The duplicated baseline and de-duplicated variant show nearly overlapping profiles across adjacent-layer pairs within the same checkpoint.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's scope as controlled data interventions, parameter-space diagnostics, external correlational validation, and a prompt de-duplication boundary case. The claims are calibrated to distinguish controlled comparisons in Section 4 from correlational external evidence in Section 5.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 8 discusses single-run intervention conditions, the limited scope of baseline run-to-run comparisons, two model scales within one architecture family, non-isolated rewriting operations, missing phase-boundary evaluations, correlational external analysis, proxy diagnostics, limited model-family coverage, and the scope of the prompt de-duplication case study.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper is empirical and diagnostic in nature and does not present formal theoretical results, theorems, or proofs.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: Section 4 specifies the model architecture, optimizer, training budget, intervention design, and checkpoint analysis; the appendix provides the rewriting prompt and additional diagnostic figures. Exact reproduction of proprietary training data may not be possible, but the intervention design and diagnostic methodology are specified sufficiently to reproduce the analysis framework.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We publicly release the data-construction and parameter-analysis code, evaluation utilities, and sample outputs at https://github.com/vivo/ScalingLens under the Apache 2.0 license. However, the proprietary training corpora and model checkpoints used in the controlled experiments cannot be released. The manuscript provides experimental configurations and detailed results to support partial reproducibility.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 specifies model size, architecture, optimizer, learning-rate schedule, batch sizes, training budget, warmup, and the definitions of Conditions A–D. The appendix provides additional details for the rewriting intervention and expanded prompt de-duplication benchmark results.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Intervention conditions have one training run each, and no confidence intervals or significance tests are reported for intervention effects. Three Condition A runs at 20k provide descriptive behavioral and spectral run-to-run comparisons, rather than uncertainty estimates for the intervention conditions.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The current manuscript specifies model scale, token budget, and training configuration, but it does not yet report GPU type, number of GPUs, wall-clock time, or approximate GPU-hours per controlled condition. Detailed GPU resource information is not included in the current manuscript and may be made available in the project repository.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work analyzes training regimes and parameter-space diagnostics and does not involve human subjects, deployment decisions, or the release of high-risk models or datasets.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

Justification: The paper is foundational research on training diagnostics and does not introduce a deployed system, new model, or new dataset. Its primary potential impact is improving transparency and monitoring of model training; no direct societal deployment or direct negative societal impact is identified.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release new pre-trained models, image generators, scraped datasets, or other high-risk assets. The external validation uses publicly released checkpoints only for analysis.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: Publicly available models, benchmarks, and software tools are cited, but their individual licenses and terms of use are not fully documented in the manuscript. The analysis code is released under Apache License 2.0 at `https://github.com/vivo/ScalingLens`. Proprietary training data and checkpoints are not publicly distributed.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The analysis and data-construction code developed for this study is publicly available at `https://github.com/vivo/ScalingLens`. The repository provides documentation, implementation scripts, and sample outputs. Proprietary training data and model checkpoints are not released.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects, so IRB approval or equivalent review is not applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

Justification: Section 4 and the appendix describe the use of a publicly available instruction model as the teacher for the Condition D rewriting intervention. This LLM usage is part of the experimental methodology rather than merely writing or editing assistance.