WIDTH SCALING IN NORMALIZED RESIDUAL NETWORKS: THE EFFECTIVE ALIGNMENT DIMENSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing theories of neural-network width characterize asymptotic limits, but provide limited guidance on whether an expansion direction identified from finite training data remains beneficial on unseen data. We study this problem for function-preserving residual expansion and introduce the effective alignment dimension, a measurable quantity describing the signal–noise geometry of activation gradients. By deriving the exact mean and variance of the inner product between independently estimated training and test gradients, we obtain a finite-sample upper bound on misalignment probability. The bound depends only on the effective alignment dimension and an effective sample size, requiring finite second moments and a nonzero population gradient, without covariance spectral assumptions or prescribed width-growth rates. We integrate this certificate into the train–test residual-expansion framework, yielding a high-probability guarantee for local finite-test directional transfer. Experiments across width-controlled LLaMA-style Transformers, Pythia, and ResNet-20 show that wider models exhibit larger effective alignment dimensions and lower empirical misalignment. Direct residual interventions confirm that the alignment statistic predicts the sign and magnitude of held-out loss changes.

1 INTRODUCTION

Width is one of the most thoroughly studied dimensions of neural scaling. A large body of theory examines large-width limits and finite-width corrections through Gaussian processes, kernel dynamics, feature-learning limits, and width-dependent fluctuations. These theories do not, however, directly provide a finite-sample certificate for whether a local expansion direction identified in a specific trained model at a specific insertion point will transfer to held-out data. This question arises naturally in model growth, where progressive stacking, Net2Net-style expansion, and learned growth operators must determine whether a proposed insertion remains beneficial beyond the finite training set.

Recent work on normalized residual networks Cheng et al. (2026) addresses this problem by decomposing depth growth into representational gain, optimization gain, and generalization transfer. It establishes a qualitative mechanism: a residual block inserted at a function-preserving zero-output initialization creates a first-order descent direction whenever its features are not orthogonal to the insertion-point error signal, while width improves the finite-sample observability of that direction. We study how reliably a training-identified direction remains favorable on independent test data.

The remaining gap is quantitative finite-sample observability. The prior width-dependent alignment result requires control of the covariance spectrum, total variance, and growth of the mean gradient signal. These assumptions yield a clean asymptotic width rate, but not an instance-specific certificate for a given model and training–test pair. Moreover, the separate worst-case conditions do not expose the relative signal–noise geometry determining finite-sample directional stability.

We retain the activation-space formulation of the prior framework. For each example, the sample-level signal is the loss gradient with respect to the hidden representation at the insertion point. Averaging over the training and test sets produces two empirical activation-gradient directions whose alignment measures whether the training-identified direction remains favorable on independent test data. Positive alignment supports the same local residual direction on both samples, whereas nonpositive alignment marks a failure of finite-sample directional transfer. Through the function-preserving jumpboard construction, this event enters the comparison between the original and expanded models.

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

Rather than controlling training and test deviations separately, we analyze their alignment statistic directly. An exact second-moment characterization yields an instance-specific finite-sample certificate summarized by a dimensionless effective alignment dimension. Together with an effective sample size determined by the training and test set sizes, it controls the probability of nonpositive alignment, with a stronger certificate as their product increases.

The result requires only a finite second moment and a nonzero population mean for the sample-level activation gradient. It requires no covariance spectral bound, prescribed width-growth rate, or fourth-moment condition, and makes no architectural assumption on the inserted module or backbone. Under the covariance-spectrum, total-variance, and signal-growth conditions of prior work, it recovers the same-order width dependence. These assumptions thus become sufficient conditions for linear growth of the effective alignment dimension with width, rather than prerequisites for the guarantee. The dimension and its signal and noise quantities can be estimated from a trained model with a small number of forward–backward passes, without explicitly forming the covariance matrix.

We then integrate the certificate into the existing direct train–test expansion framework. Under the original assumptions ensuring an improving jumpboard model, realizability of the residual direction, subsequent optimization gain, and uniform generalization control, the resulting theorem replaces the previous alignment-control step with an instance-specific guarantee in terms of the training and test set sizes and the effective alignment dimension, while leaving the representational, optimization, and generalization components unchanged. The probability of successful directional transfer is therefore expressed directly through these quantities.

This analysis also clarifies the role of width. Width does not enter the certificate independently, but changes the activation-gradient distribution and hence the relative signal–noise geometry summarized by the effective alignment dimension. If the relative-noise structure does not deteriorate with width, increasing width cannot worsen the certificate; if the effective alignment dimension increases, the alignment-failure bound decreases. The benefit of width is therefore conditional rather than automatic and can be evaluated on the model at hand.

Experiments on width-controlled LLaMA-style Transformers, Pythia, and ResNet-20 support this account: d_{align} generally increases with width while empirical misalignment decreases under fixed sampling budgets. Varying the training and test sample sizes further supports the predicted finite-sample dependence. Direct residual interventions verify that positive alignment predicts the sign and local magnitude of realizable held-out loss changes.

Contributions. ❶ We derive an exact second-moment characterization of the alignment between training- and test-averaged activation gradients, yielding an instance-specific finite-sample certificate summarized by the effective alignment dimension. ❷ We replace the previous width-dependent alignment step in the direct train–test expansion framework with this certificate, obtaining an instance-specific finite-sample guarantee for directional transfer in terms of the training and test sample sizes and the effective alignment dimension. ❸ Across width-controlled Transformers and ResNets, we show that the effective alignment dimension generally increases with width alongside reduced empirical misalignment. Direct residual interventions further verify that the alignment statistic predicts both the sign and local magnitude of held-out loss changes.

2 Related Work

Theoretical studies of neural-network width primarily characterize large-width limits and deviations around them. NNGP and NTK limits yield Gaussian-process and kernel descriptions (Neal, 1995; Lee et al., 2018; Jacot et al., 2018), while mean-field and maximal-update parameterizations retain nontrivial feature learning (Mei et al., 2018; Chizat & Bach, 2018; Yang & Hu, 2021; Yang et al., 2021). Finite-width analyses quantify width-dependent fluctuations of kernels and predictions (Bordelon & Pehlevan, 2023), whereas neural scaling-law theories describe average loss as data or model size grows (Bahri et al., 2021). A separate literature studies coherent gradients, parameter-gradient signal-to-noise ratios, and gradient noise scale as diagnostics of optimization and generalization (Chatterjee, 2020; Liu et al., 2020; McCandlish et al., 2018), but does not analyze the two-sample sign-reversal probability of an update selected for a newly inserted module. Function-preserving model-growth methods, including Net2Net, progressive stacking, bert2BERT, and LiGO, reuse trained models to initialize larger ones (Chen et al., 2015; Gong et al., 2019; Chen et al., 2022; Wang et al., 2023);

concurrent work emphasizes the geometry of function preservation and post-insertion plasticity (Lok, 2026; Khemais, 2026). Closest to our setting, (Cheng et al., 2026) decompose residual expansion into representational, optimization, and generalization components and derive a width-dependent train/test activation-gradient alignment guarantee.

3 Assumptions and Notation

Appendix G and Appendix H provide detailed notation, a concrete residual-insertion example, and a self-contained account of the relevant prior results and proof chain. We adopt the activation-space notation and direct train–test comparison framework of (Cheng et al., 2026). We write q for the insertion-point activation gradient, denoted ζ in their train–test alignment theorem.

Throughout the paper, we fix a well-trained reference model f^*_{old}, and $f_{\mathrm{old}} = f_{\mathrm{top}} \circ f_{\mathrm{bot}} \in \mathcal{H}_{\mathrm{old}}$ and a candidate insertion point. For an input x, let $z(x) := f_{\mathrm{bot}}(x) \in \mathbb{R}^N$ denote the hidden representation at that point, where N is its width. Let $\mathcal{H}_{\mathrm{new}}$ denote the hypothesis class obtained by inserting a residual function at the selected location. For any predictor f and loss function ℓ, define the population, training, and test risks by $R(f) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell\big(f(x),y\big)\right]$, $\mathcal{L}_{\mathrm{train}}(f) = \frac{1}{M}\sum_{i=1}^{M}\ell\big(f(x_i),y_i\big)$, $\mathcal{L}_{\mathrm{test}}(f) := \frac{1}{K}\sum_{j=1}^{K}\ell\big(f(\widetilde{x}_j),\widetilde{y}_j\big)$.

3.1 Basic Assumptions

Assumption 1 (Independent sampling and second moments)**.** *Let $\mathcal{S}_{\mathrm{train}} = \{(x_i,y_i)\}_{i=1}^{M}$, $\mathcal{S}_{\mathrm{test}} = \{(\widetilde{x}_j,\widetilde{y}_j)\}_{j=1}^{K}$ be independent samples satisfying $\mathcal{S}_{\mathrm{train}} \overset{\text{i.i.d.}}{\sim} \mathcal{D}^M$, $\mathcal{S}_{\mathrm{test}} \overset{\text{i.i.d.}}{\sim} \mathcal{D}^K$, $\mathcal{S}_{\mathrm{train}} \perp \mathcal{S}_{\mathrm{test}}$.*

All alignment probabilities are understood conditional on the fixed reference model f^*_{old} and the fixed insertion point.

For a sample (x,y), define the activation-gradient signal at the insertion point by $q(x,y) := \nabla_z \ell\big(f_{\mathrm{top}}(z),y\big)\big|_{z=z(x)} \in \mathbb{R}^N$.

Define its population mean and covariance by $\bar{\mu} := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(x,y)\right]$, $\Sigma := \mathrm{Cov}_{(x,y)\sim\mathcal{D}}\left(q(x,y)\right)$. We assume $\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\|q(x,y)\|_2^2\right] < \infty, \bar{\mu} \neq 0$.

Assumption 1 is the only assumption required for Theorem 1. In particular, the finite-sample alignment result requires no covariance spectral bound, fourth-moment condition, or prescribed width-growth rate.

Assumption 2 (Selection condition)**.** *Let $\widetilde{f}_S \in \mathcal{H}_{\mathrm{new}}$ denote the empirical jumpboard model constructed from $\mathcal{S}_{\mathrm{train}}$ in the direct train–test expansion framework, and let $f_{\mathrm{new}} \in \mathcal{H}_{\mathrm{new}}$ denote the final expanded model.*

We assume $\mathcal{L}_{\mathrm{train}}\left(f_{\mathrm{new}}\right) \le \mathcal{L}_{\mathrm{train}}\left(\widetilde{f}_S\right)$. Equivalently, define the nonnegative optimization gain by $\Delta_{\mathrm{ERM}} := \mathcal{L}_{\mathrm{train}}\left(\widetilde{f}_S\right) - \mathcal{L}_{\mathrm{train}}\left(f_{\mathrm{new}}\right) \ge 0$.

Assumption 3 (Uniform generalization control)**.** *Fix a confidence parameter $\delta \in (0,1)$. We assume that there exist deterministic generalization radii $\epsilon_M, \epsilon_K \ge 0$ such that $\Pr\left[\sup_{f\in\mathcal{H}_{\mathrm{new}}} |R(f) - \mathcal{L}_{\mathrm{train}}(f)| \le \epsilon_M\right] \ge 1-\delta$ and $\Pr\left[\sup_{f\in\mathcal{H}_{\mathrm{new}}} |R(f) - \mathcal{L}_{\mathrm{test}}(f)| \le \epsilon_K\right] \ge 1-\delta$.*

The boundedness, Lipschitz, normalization, and norm-control conditions used by Cheng et al. (2026) to instantiate ϵ_M and ϵ_K are restated in the appendix. The main text uses only the two uniform inequalities in Assumption 3.

For the direct train–test comparison, define the finite-test jumpboard margin by $\Delta_R^{\mathrm{test}} := \mathcal{L}_{\mathrm{test}}\left(f^*_{\mathrm{old}}\right) - \mathcal{L}_{\mathrm{test}}\left(\widetilde{f}_S\right)$.

Despite the subscript R inherited from the prior notation, Δ_R^{test} denotes a finite-test empirical margin rather than a population-risk difference.

The small-step and residual-realizability conditions used to convert positive activation-gradient alignment into a positive finite-test jumpboard margin are invoked directly from Cheng et al. (2026, Corollary 3) and restated in the appendix. These conditions apply only to the additional margin statement in Theorem 2; they are not assumptions of Theorem 1 and are therefore not repeated as global assumptions here.

3.2 Key Alignment Quantities

Define the empirical activation-gradient averages over the training and test sets by $\mu_M := \frac{1}{M}\sum_{i=1}^{M} q(x_i, y_i)$ and $g_K := \frac{1}{K}\sum_{j=1}^{K} q(\widetilde{x}_j, \widetilde{y}_j)$. The finite-sample activation-gradient misalignment event is $\{\mu_M^\top g_K \le 0\}$.

Define the directional effective dimension as $d_\parallel := \frac{\|\bar{\mu}\|_2^4}{\bar{\mu}^\top \Sigma \bar{\mu}}$, the covariance-energy effective dimension as $d_2 := \frac{\|\bar{\mu}\|_2^4}{\mathrm{tr}(\Sigma^2)} = \frac{\|\bar{\mu}\|_2^4}{\|\Sigma\|_F^2}$, and the effective alignment dimension as $d_{\mathrm{align}} := \min\{d_\parallel, d_2\}$. If a denominator is zero, the corresponding effective dimension is defined as $+\infty$.

The exact certificate uses $\Xi_{M,K} := \frac{1}{Md_\parallel} + \frac{1}{Kd_\parallel} + \frac{1}{MKd_2}$ and $B_{\mathrm{exact}}(M,K) := \frac{\Xi_{M,K}}{1+\Xi_{M,K}}$. With the effective sample-size factor $n_{\mathrm{eff}}(M,K) := \frac{MK}{M+K+1}$, the simplified certificate is $B_{\mathrm{align}}(M,K) := \frac{1}{1+n_{\mathrm{eff}}(M,K)d_{\mathrm{align}}}$. The two certificates satisfy $B_{\mathrm{exact}}(M,K) \le B_{\mathrm{align}}(M,K) < 1$.

When dependence on insertion width must be explicit, we write $\bar{\mu}^{(N)}$, $\Sigma^{(N)}$, $d_\parallel(N)$, $d_2(N)$, and $d_{\mathrm{align}}(N)$. At a fixed insertion width, the width argument is omitted.

Theorem 1 uses only Assumption 1. The core risk comparison in Theorem 2 uses Assumptions 2 and 3. Its additional positive-margin statement combines Theorem 1 with the small-step and residual-realizability conditions of Cheng et al. (2026, Corollary 3), which are restated in the appendix.

4 Main Results

4.1 Main Theorems

Our first result controls the probability that the activation-gradient direction estimated from the training set fails to remain positively aligned on the independent test set. In addition to the probability bound, we state the exact mean and variance identities that generate the two effective dimensions $d_\parallel$ and d_2.

Theorem 1 (Finite-sample activation-gradient alignment)**.** *Under Assumption 1, the train–test alignment statistic satisfies* $\mathbb{E}\left[\mu_M^\top g_K\right] = \|\bar{\mu}\|_2^2$ *and* $\mathrm{Var}\left(\mu_M^\top g_K\right) = \left(\frac{1}{M} + \frac{1}{K}\right)\bar{\mu}^\top\Sigma\bar{\mu} + \frac{1}{MK}\,\mathrm{tr}\left(\Sigma^2\right)$. *Consequently,*

$$\Pr\left(\mu_M^\top g_K \le 0\right) \le B_{\mathrm{exact}}(M,K) = \frac{\Xi_{M,K}}{1+\Xi_{M,K}},$$

Moreover, the one-scalar certificate satisfies

$$\Pr\left(\mu_M^\top g_K \le 0\right) \le B_{\mathrm{exact}}(M,K) \le B_{\mathrm{align}}(M,K) = \frac{1}{1+n_{\mathrm{eff}}(M,K)d_{\mathrm{align}}}$$

The exact certificate retains the two relative-noise contributions separately. The simplified certificate replaces them by their conservative minimum, the effective alignment dimension. The theorem requires neither a covariance spectral bound nor a prescribed width-growth law: those conditions are unnecessary for the fixed-width finite-sample guarantee.

The next result combines the new alignment certificate with the direct train–test comparison framework. Its first part is the algebraic comparison between the original model, the empirical jumpboard, and the final expanded model. Its second part replaces the previous width-dependent alignment bound by applying Theorem 1 to the existence of a locally improving function-preserving path.

Theorem 2 (Direct train–test expansion with effective-alignment control)**.** *Under Assumptions 2 and 3, with probability at least* $1-2\delta$*, the final expanded model satisfies*

$$\mathcal{L}_{\mathrm{test}}\left(f_{\mathrm{new}}\right) \le \mathcal{L}_{\mathrm{test}}\left(f^*_{\mathrm{old}}\right) - \Delta_R^{\mathrm{test}} - \Delta_{\mathrm{ERM}} + 2\left(\epsilon_M + \epsilon_K\right).$$

Consequently, on every realization for which $\Delta_R^{\text{test}} + \Delta_{\text{ERM}} > 2\,(\epsilon_M + \epsilon_K)$*, one has the strict test-risk improvement* $\mathcal{L}_{\text{test}}\,(f_{\text{new}}) < \mathcal{L}_{\text{test}}\,(f^*_{\text{old}})$.

Suppose additionally that the empirical jumpboard is chosen in the small-step regime and satisfies the residual-realizability conditions of Cheng et al. (2026, Corollary 3). Then positive activation-gradient alignment implies a positive finite-test jumpboard margin. Hence, by Theorem 1,

$$\Pr\left(\Delta_R^{\text{test}} > 0\right) \geq 1 - B_{\text{exact}}(M, K) \geq 1 - B_{\text{align}}(M, K).$$

The two uniform-generalization events and the positive-margin event therefore hold simultaneously with probability at least $1 - 2\delta - B_{\text{exact}}(M, K)$*, and using the simplified certificate, with probability at least* $1 - 2\delta - B_{\text{align}}(M, K)$.

The first part of Theorem 2 does not require the alignment event and remains valid even when the finite-test margin is zero or negative. The alignment-supported part instead quantifies how often this random margin is positive. Importantly, Theorem 1 controls the sign of the margin through the inherited jumpboard bridge; it does not lower-bound the magnitude of the margin. Strict improvement therefore continues to require the realized margin and optimization gain to dominate the two generalization radii.

4.2 Width Dependence of the Alignment Certificate

Theorem 1 is a fixed-width finite-sample result. To compare models of different widths, we make the width dependence of the activation-gradient distribution and the effective alignment dimension explicit. The simplified certificate becomes $B_{\text{align}}(M, K; N)$.

This factorization separates the effects of data and model width. The effective sample-size factor depends only on the training and test set sizes. Width instead changes the distribution of the activation gradient, and hence its population mean, covariance, and effective alignment dimension. Width therefore does not improve the certificate directly; it does so only when it improves the relative signal–noise geometry summarized by $d_{\text{align}}(N)$.

A sufficient condition for a wider model to have a non-worse simplified certificate is that neither component of the effective alignment dimension deteriorates. In particular, if $N_2 > N_1$, $d_{\|}(N_2) \geq d_{\|}(N_1)$ and $d_2(N_2) \geq d_2(N_1)$, then $B_{\text{align}}(M, K; N_2) \leq B_{\text{align}}(M, K; N_1)$. This is a sufficient rather than necessary condition: the exact certificate in Theorem 1 retains the two relative-noise contributions separately and may give a finer comparison.

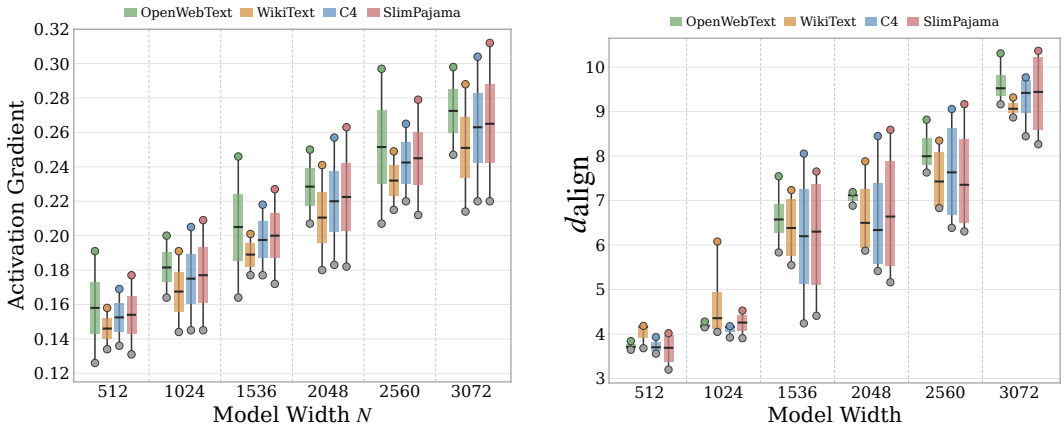
The stronger covariance and signal-growth conditions used in the prior width-dependent analysis provide one sufficient route to this behavior. Indeed, the bounds in Appendix J give, for sufficiently large N, $d_{\|}(N) \geq \|\bar{\mu}(N)\|_2^2/\lambda_{\max}(\Sigma(N)) \geq (c/C)N$ and $d_2(N) \geq \|\bar{\mu}(N)\|_2^4/[\lambda_{\max}(\Sigma(N))\,\text{tr}(\Sigma(N))] \geq [c^2/(CC_{\text{tr}})]N$, where $c, C, C_{\text{tr}} > 0$ are independent of N. Hence $d_{\text{align}}(N) \geq c_0 N$ for $c_0 = \min\{c/C, c^2/(CC_{\text{tr}})\} > 0$. Consequently,

$$\Pr(\mu_M^\top g_K \leq 0) \leq \frac{1}{1 + c_0 n_{\text{eff}}(M, K)N} = O\left(\frac{1}{n_{\text{eff}}(M, K)}\frac{1}{N}\right).$$

Up to the lower-order mixed term, this recovers the width dependence of the prior alignment result. The difference is that these growth conditions are not required by Theorem 1. They are only sufficient conditions for linear growth of the effective alignment dimension. More generally, the certificate improves at whatever rate is supported by the observed behavior of $d_{\text{align}}(N)$. Under the local conditions of Theorem 2, the same bound also controls failure of a positive finite-test jumpboard margin. Thus, the benefit of width for finite-sample directional transfer is conditional rather than automatic and can be evaluated directly on the model at hand.

5 Experiment

Our experiments evaluate four questions: how activation-gradient geometry scales with width, whether empirical misalignment follows the predicted width and sample-size dependence, whether the trends transfer across Pythia and ResNet, and whether positive alignment predicts the loss change of a realizable residual intervention.



(a) Activation-gradient statistic across model widths.

(b) d_{align} across model widths.

Figure 1: Width-dependent activation-gradient geometry. Panel (a) reports the activation-gradient statistic $\widehat{S} = \|\widehat{\mu}\|_2^2$, whereas panel (b) reports d_{align} computed from the same activation-gradient samples. For each width, metric estimates are first aggregated across independently sampled evaluation subsets within each trained model and then summarized across eight training seeds.

5.1 Experiment Configurations

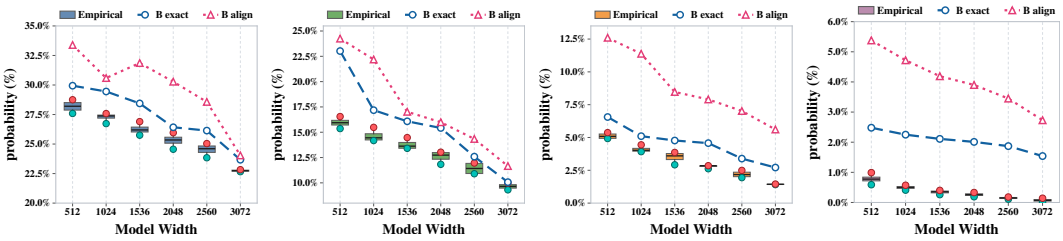
Model. We consider three experimental settings: our pretrained 32-layer LLaMA-style Transformer models (Touvron et al., 2023), the Pythia model suite, and independently trained ResNet models. For the LLaMA-style family, we pretrain six models with hidden widths $N \in \{512, 1024, 1536, 2048, 2560, 3072\}$ while keeping the depth, training data, token budget, and optimization settings fixed. For each LLaMA-style and ResNet configuration, we train eight independent models using different random seeds. We further evaluate Pythia checkpoints of different scales.

Metric Evaluation. For LLaMA-style and Pythia models, we measure the activation gradient q at the final Transformer block, and for ResNet-20 at the first residual block. For each language-modeling sequence (x_i, y_i), we define the sequence-level signal using a token-shared additive residual b as $q_i = \nabla_b \ell(\{z_{i,t} + b\}_{t=1}^T, y_i)|_{b=0} = \sum_{t=1}^T \nabla_{z_{i,t}} \ell$, where b where token-level normalization is included in the sequence loss ℓ. We estimate d_{align} and the empirical misalignment probability from independently sampled, non-overlapping activation-gradient subsets, using multiple data-sampling seeds on C4, WikiText-2, OpenWebText, and SlimPajama for language models, and CIFAR-10/100 for ResNet-20. For independently trained models, sampling variability is first aggregated within each checkpoint and then summarized across training seeds. Appendix M motivates the layer choices and reports cross-layer $d_{\|}$ and d_2 for LLaMA-style models; the theoretical existence claim requires only one admissible insertion layer to satisfy the stated conditions.

5.2 Width Scaling of Activation-Gradient Geometry

We first examine how activation-gradient geometry varies with model width in the controlled LLaMA-style family. We consider hidden widths $N \in \{512, 1024, 1536, 2048, 2560, 3072\}$ while holding the model depth, training corpus, token budget, and optimization protocol fixed. Figure 1 reports the activation-gradient statistic $\widehat{S}$ and the effective alignment dimension d_{align}, both computed from the same activation-gradient samples, on C4, WikiText-2, OpenWebText, and SlimPajama. Detailed metric-estimation procedures and pseudocode are provided in the appendix.

As shown in Figure 1a, the activation-gradient statistic exhibits a clear width-dependent trend. Although its absolute scale varies across datasets and individual runs, the distributions generally shift upward as N increases. This pattern persists across independently trained models and independently sampled evaluation subsets, indicating that the observed trend is robust to both optimization randomness and finite-sample estimation variability.



(a) $M = K = 1$. (b) $M = K = 2$. (c) $M = K = 4$. (d) $M = K = 8$.

Figure 2: Empirical gradient-misalignment probabilities and empirical estimates across model widths and finite-sample budgets. Each sampled sequence contains 2048 tokens, and each empirical probability is estimated using $R = 500{,}000$ independent Monte Carlo trials. The empirical estimates $\widehat{B}_{\text{exact}}$ and $\widehat{B}_{\text{align}}$ are computed from the corresponding activation-gradient moments. Results are summarized across eight independently trained LLaMA-style models and multiple independent data-sampling seeds.

Figure 1b shows a corresponding increase in the effective alignment dimension computed from the activation-gradient signal q. Across the four datasets, d_{align} grows from approximately 3–4 at the smallest widths to approximately 9–10 at the largest width. Although adjacent widths exhibit overlapping distributions and moderate local fluctuations, the aggregated trend remains consistently increasing over the investigated width range. Taken together, Figure 1 provides empirical support for the central width-scaling hypothesis

$$\boldsymbol{N \uparrow \implies d_{\text{align}}(N) \uparrow}$$

5.3 Gradient-Misalignment Probability

We next directly test the theoretical prediction that finite-sample gradient misalignment becomes less likely as the effective alignment dimension increases. To isolate the effect of model width from variation in the evaluation distribution, we conduct this analysis exclusively on C4 using our trained LLaMA-style models. For each trained model and sampling budget (M, K), we estimate the gradient-misalignment probability using the Monte Carlo estimator

$$\widehat{p}_{M,K}(N) = \frac{1}{R} \sum_{r=1}^{R} \mathbf{1}\left\{\widehat{\mu}_{M,r}^{\top} \widehat{g}_{K,r} \le 0\right\}, \tag{1}$$

where $\widehat{\mu}_{M,r}$ and $\widehat{g}_{K,r}$ are computed from two independently sampled subsets containing M and K sequences, respectively, in Monte Carlo trial r. Each sampled sequence contains 2048 tokens. Independently sampling the two subsets matches the independence condition in Theorem 1 and avoids introducing artificial correlation between the two finite-sample gradient estimates.

We evaluate $M = K \in \{1, 2, 4, 8\}$ and perform $R = 500{,}000$ conditionally independent Monte Carlo trials for each combination of model width, training seed, and sampling budget. These Monte Carlo trials are repeated resamples from fixed gradient pools and are distinct from the independent training runs and data-sampling seeds used to quantify outer variability. We additionally report the empirical estimates $\widehat{B}_{\text{exact}}$ and $\widehat{B}_{\text{align}}$, computed by substituting empirical activation-gradient moments into the corresponding population bounds in Theorem 1.

Figure 2 shows that the empirical gradient-misalignment probability consistently decreases as model width increases, and that this trend holds for every fixed sampling budget. Because M and K remain constant within each panel, the observed reduction cannot be attributed to wider models receiving more evaluation data. Instead, wider models yield more statistically reliable finite-sample gradient directions under the same sampling budget.

Both empirical estimates exhibit the same decreasing trend with width and remain conservative relative to the empirical probabilities. As expected from Theorem 1, $\widehat{B}_{\text{exact}}$ is consistently tighter than $\widehat{B}_{\text{align}}$, while the latter provides a simpler one-scalar characterization through d_{align}. The

agreement in trend between the empirical probabilities and the two estimates supports the predicted connection between activation-gradient geometry and finite-sample directional transfer.

For $M = K = 1$, the empirical misalignment probability decreases from approximately 28.5% at $N = 512$ to approximately 22.7% at $N = 3072$. For $M = K = 2$, it decreases from approximately 15.7% to approximately 9.7%. At larger sampling budgets, the reduction becomes more pronounced relative to the initial probability: for $M = K = 4$, the probability decreases from approximately 5.2% to approximately 1.4%, whereas for $M = K = 8$, it decreases from approximately 0.7% to below 0.1%.

At every tested width, misalignment falls sharply as $M = K$ increases, especially from 1 to 8, and is nearly zero above 16. Thus, model width and sample size improve directional reliability through distinct mechanisms: increasing width improves the underlying activation-gradient geometry, whereas increasing M and K reduces finite-sample estimation uncertainty. Together with the width-dependent increase in d_{align} observed in Figure 1, these results provide empirical support for the mechanism

$$N \uparrow \implies d_{\text{align}}(N) \uparrow \implies p_{M,K}(N) \downarrow$$

5.4 Cross-Architecture and Cross-Modality Validation

5.4.1 Fixed-Depth Width Validation on Pythia

We next evaluate two fixed-depth Pythia regimes to separate width from depth. The six-layer regime contains Pythia-14M, 31M, and 70M with widths $N \in \{128, 256, 512\}$ and ten training seeds per scale; the 32-layer regime contains Pythia-2.8B and 6.9B with widths $N \in \{2560, 4096\}$. Both $\widehat{S}$ and d_{align} are computed from the same activation-gradient samples.

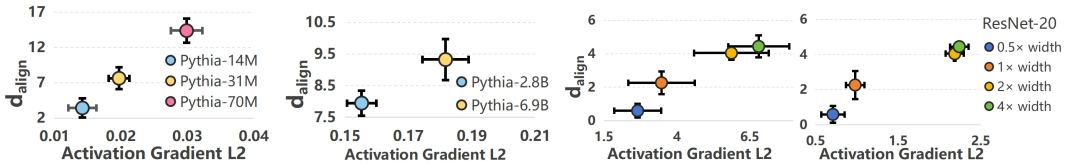
To disentangle the effect of width from that of depth, we separately analyze two fixed-depth Pythia regimes. The small-scale group consists of Pythia-14M, Pythia-31M, and Pythia-70M, all with six Transformer layers and hidden widths $N \in \{128, 256, 512\}$. For each model scale, we evaluate ten independently trained runs corresponding to distinct training seeds, including the additional seed replicas released by PolyPythias (van der Wal et al., 2025). The large-scale group consists of Pythia-2.8B and Pythia-6.9B, both with 32 Transformer layers and hidden widths $N \in \{2560, 4096\}$. Within each group, model depth is fixed while hidden width and the architectural dimensions conventionally coupled to width increase.

Figure 3a shows a clear monotonic progression within the six-layer regime. As the hidden width increases from 128 to 512, the activation-gradient statistic $\widehat{S}$ increases from approximately 0.014 to 0.030, while d_{align} increases from approximately 3 to 14. Thus, the three model configurations move consistently upward and to the right in the gradient-geometry plane, reproducing the width-dependent behavior observed in our controlled LLaMA-style models. Figure 3b provides a complementary validation at substantially larger scale. With depth fixed at 32 layers, increasing the hidden width from 2560 in Pythia-2.8B to 4096 in Pythia-6.9B increases both the activation-gradient statistic $\widehat{S}$ and d_{align}.

5.4.2 Cross-Architecture Validation on ResNet

We next test whether the observed width-dependent gradient geometry extends beyond autoregressive Transformers to convolutional vision models. We train ResNet-20 models (He et al., 2016) from scratch on CIFAR-10 and CIFAR-100 (Krizhevsky & Hinton, 2009) with channel-width multipliers $\{0.5\times, 1\times, 2\times, 4\times\}$. Within each dataset, network depth and all training conditions are held fixed, while only the channel dimensions and their associated width-dependent parameters are scaled. Each configuration is independently trained with eight random seeds, and the proposed metrics are evaluated using multiple data-sampling seeds. Metric estimates are first aggregated across data-sampling seeds within each trained model and are then summarized across the eight independent training seeds. Optimization hyperparameters and training schedules are provided in the appendix.

Figure 3c shows a consistent progression in activation-gradient geometry as channel width increases. On both CIFAR-10 and CIFAR-100, wider ResNet models move upward and to the right in the gradient-geometry plane, indicating increases in both the activation-gradient statistic $\widehat{S}$ and d_{align}.



(a) Six-layer Pythia. (b) 32-layer Pythia. (c) ResNet-20 on CIFAR-10 and CIFAR-100.

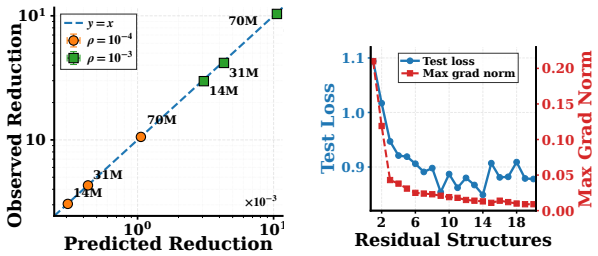
Figure 3: Cross-architecture width validation on fixed-depth Pythia models and ResNet-20 on CIFAR-10 (left) and CIFAR-100 (right). Wider models generally exhibit stronger activation-gradient signals and larger d_{align}; error bars summarize variability across runs or evaluation subsets.

This trend remains stable across independently trained models and evaluation samples despite differences in dataset complexity and metric scale.

Increasing the width multiplier from $0.5\times$ to $4\times$ produces substantial gains in both quantities on CIFAR-10, with the same qualitative trend observed on CIFAR-100. The smaller difference between the $2\times$ and $4\times$ settings indicates a weaker marginal increase over the investigated range.

5.5 Empirical Validation of Theorem 2

We validate the local alignment-to-path implication of Theorem 2 through direct residual interventions. A zero-initialized token-shared residual parameter is inserted before the final normalization layer of Pythia-14M, 31M, and 70M, using ten independently trained checkpoints per scale. In each C4 trial, a training subset determines μ_M and the residual direction, while an independent held-out subset provides g_K. Thus, the sign of $\mu_M^\top g_K$ predicts whether the selected direction decreases held-out loss, with the norm-matched reverse direction giving the opposite sign. Across 15,000 trials, the observed loss changes follow these predictions. Figure 4(a) also shows close agreement between predicted and observed reductions for $\rho \in \{10^{-4}, 10^{-3}\}$. We then study the same mechanism during model growth by sequentially adding residual structures to a ResNet on CIFAR-100. Figure 4(b) shows that test loss decreases rapidly before entering a flatter regime, while the maximum activation-gradient norm drops from about 0.21 to 0.01. These results support the small-step positive-margin statement in Theorem 2: positive train–test alignment identifies an improving realizable jumpboard in the local regime, while the finite-step interventions separately demonstrate successful selected steps in the tested regime. Repeated expansion gives diminishing gains as the available signal weakens.



(a) Direct intervention. (b) Sequential expansion.

Figure 4: Empirical validation of Theorem 2. (a) Predicted versus observed held-out loss reductions. (b) Test loss and maximum activation-gradient norm during sequential residual expansion.

6 Conclusion

We studied when a local expansion direction identified from finite training data transfers to independent test data. Using the insertion-point activation gradient, we derived an exact finite-sample misalignment bound controlled by the effective sample size and d_{align}. Width enters the certificate through the activation-gradient distribution: whenever $d_{\text{align}}(N)$ increases with width, the misalignment bound decreases. Under the stronger covariance and signal-growth conditions of prior work, the certificate recovers the same-order width dependence. We incorporated this certificate into the direct train–test expansion framework. Experiments on width-controlled LLaMA-style Transformers, Pythia, and ResNet-20 show that d_{align} generally increases with width while empirical misalignment decreases under fixed sampling budgets. Direct residual interventions further confirm that positive alignment predicts the sign and local magnitude of held-out loss changes.

AI USE STATEMENT

We used GPT-6 solely for language polishing. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final paper.

REFERENCES

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*, 2021. URL https://arxiv.org/abs/2102.06701.

Blake Bordelon and Cengiz Pehlevan. Dynamics of finite width kernel and prediction fluctuations in mean field neural networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1ec69275e9f002ee068f5d68380f3290-Abstract-Conference.html.

Satrajit Chatterjee. Coherent gradients: An approach to understanding generalization in gradient descent-based optimization. *arXiv preprint arXiv:2002.10657*, 2020. URL https://arxiv.org/abs/2002.10657.

Cheng Chen, Yichun Yin, Lifeng Shang, Xin Jiang, Yujia Qin, Fengyu Wang, Zhi Wang, Xiao Chen, Zhiyuan Liu, and Qun Liu. bert2BERT: Towards reusable pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2134–2148. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.151. URL https://aclanthology.org/2022.acl-long.151/.

Tianqi Chen, Ian Goodfellow, and Jonathon Shlens. Net2Net: Accelerating learning via knowledge transfer. *arXiv preprint arXiv:1511.05641*, 2015. URL https://arxiv.org/abs/1511.05641.

Daning Cheng, Zeyu Liu, Jun Sun, Fen Xia, Boyang Zhang, Dongping Liu, and Yunquan Zhang. A qualitative test-risk mechanism for scaling behavior in normalized residual networks. *arXiv preprint arXiv:2605.08297*, 2026. URL https://arxiv.org/abs/2605.08297.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/a1afc58c6ca9540d057299ec3016d726-Abstract.html.

Linyuan Gong, Di He, Zhuohan Li, Tao Qin, Liwei Wang, and Tieyan Liu. Efficient training of BERT by progressively stacking. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2337–2346. PMLR, 2019. URL https://proceedings.mlr.press/v97/gong19a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html.

Abdallah Khemais. Exact network surgery: Functional invariance and gradient plasticity in reactive computational graphs. *arXiv preprint arXiv:2607.16568*, 2026. URL https://arxiv.org/abs/2607.16568.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1EA-M-0Z.

Jinlong Liu, Yunzhi Bai, Guoqing Jiang, Ting Chen, and Huayan Wang. Understanding why neural networks generalize well through GSNR of parameters. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HyevIJStwH.

Dante Lok. Gate-zero growth: A geometric framework for function-preserving continual learning. *arXiv preprint arXiv:2607.14571*, 2026. URL https://arxiv.org/abs/2607.14571.

Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training. *arXiv preprint arXiv:1812.06162*, 2018. URL https://arxiv.org/abs/1812.06162.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018. doi: 10.1073/pnas.1806579115.

Radford M. Neal. *Bayesian Learning for Neural Networks*. PhD thesis, University of Toronto, 1995.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. URL https://arxiv.org/abs/2302.13971.

Oskar van der Wal, Pietro Lesci, Max Müller-Eberstein, Naomi Saphra, Hailey Schoelkopf, Willem H. Zuidema, and Stella Biderman. Polypythias: Stability and outliers across fifty language model pre-training runs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bmrYu2Ekdz.

Peihao Wang, Rameswar Panda, Lucas Torroba Hennigen, Philip Greengard, Leonid Karlinsky, Rogerio Feris, David Daniel Cox, Zhangyang Wang, and Yoon Kim. Learning to grow pretrained models for efficient transformer training. *arXiv preprint arXiv:2303.00980*, 2023. URL https://arxiv.org/abs/2303.00980.

Greg Yang and Edward J. Hu. Tensor programs IV: Feature learning in infinite-width neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 11727–11737. PMLR, 2021. URL https://proceedings.mlr.press/v139/yang21c.html.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17084–17097, 2021. URL https://openreview.net/forum?id=Bx6qKuBM2AD.

Appendix

1 Introduction . . . 1

2 Related Work . . . 2

3 Assumptions and Notation . . . 3

3.1 Basic Assumptions . . . 3

3.2 Key Alignment Quantities . . . 4

4 Main Results . . . 4

4.1 Main Theorems . . . 4

4.2 The Role of Width . . . 5

5 Experiment . . . 5

5.1 Experiment Configurations . . . 6

5.2 Width Scaling of Activation-Gradient Geometry . . . 6

5.3 Gradient-Misalignment Probability . . . 7

5.4 Cross-Architecture and Cross-Modality Validation . . . 8

5.5 Empirical Validation of Theorem 2 . . . 9

6 Conclusion . . . 9

A Model Architecture and Pretraining Details . . . 13

A.1 Model Architecture . . . 13

A.2 Model Configurations . . . 13

A.3 Pretraining Data . . . 14

A.4 Pretraining Protocol . . . 14

A.5 Independent Training Runs . . . 15

B Training Dynamics and Optimization Stability . . . 15

B.1 Training Loss . . . 15

B.2 Validation Loss and Late-Training Perplexity . . . 16

B.3 Gradient Norm and Training Stability . . . 16

B.4 Pythia Model Configurations . . . 17

B.5 ResNet-20 Experimental Configuration and Accuracy . . . 18

C Computation of Activation-Gradient Alignment Statistics . . . 19

D Monte Carlo Estimation of Misalignment Probability . . . 20

E Component-wise Decomposition of the Effective Alignment Dimension . . . 21

F Evolution of Gradient-Misalignment Probability during Pretraining . . . 22

G Problem Setup and Notation . . . 26

H Residual Expansion Setup and Connection to Prior Work . . . 29

I Proof of Theorem 1 . . . 32

J Proof of Theorem 2 . . . 34

K Recovery of the Previous Width-Dependent Rate . . . 37

L Exact Parameter-Space Realization of the Activation-Gradient Direction . . . 39

M Architecture-Specific Selection of Activation-Gradient Measurement Locations . . . 40

M.1 Existential Nature of the Layer-Wise Theory . . . 41

M.2 Why the Final Residual Representation Is Measured in LLMs . . . 41

M.3 Why an Early Residual Block Is Measured in ResNet . . . 43

M.4 Layer-Wise Component Decomposition of the Effective Alignment Dimension . . . 46

A Model Architecture and Pretraining Details

A.1 Model Architecture

All models use the same Llama-style decoder-only Transformer architecture. Each Transformer block adopts a pre-normalization structure consisting of RMSNorm, grouped-query self-attention with rotary positional embeddings (RoPE), a SwiGLU feed-forward network, and residual connections around both sublayers. A final RMSNorm and a linear language-modeling head map the hidden representations to next-token logits. We use the same tokenizer and vocabulary for all model configurations and do not apply supervised fine-tuning, RLHF, or any other post-training procedure.

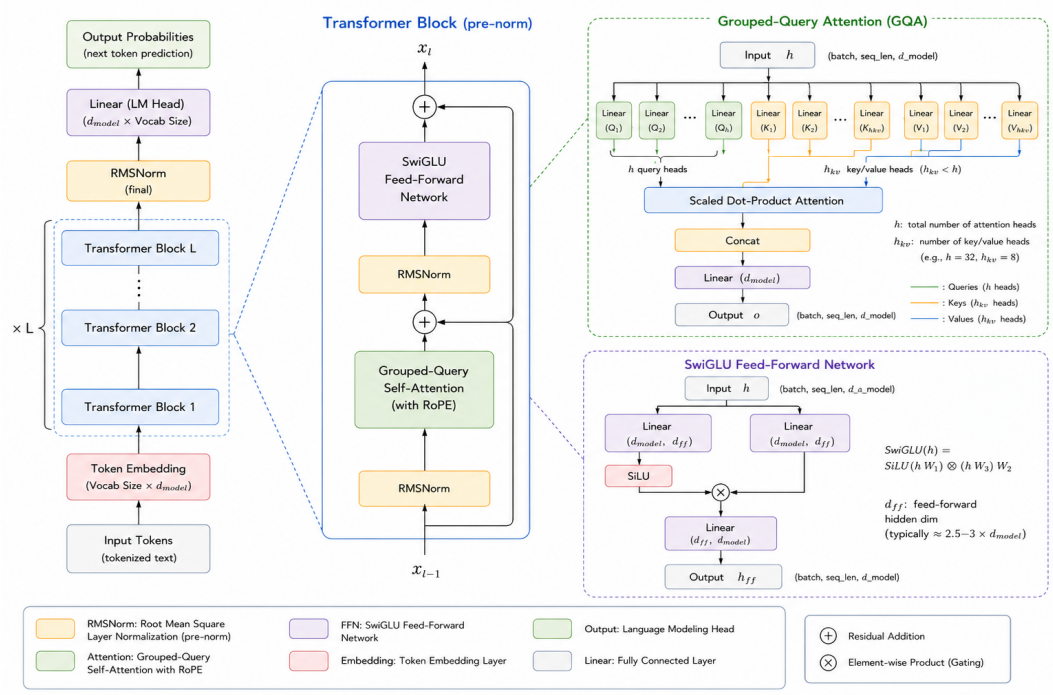


Figure 5: Decoder-only Transformer architecture shared by all model configurations. Each pre-normalized block contains grouped-query self-attention with rotary positional embeddings, followed by a SwiGLU feed-forward network; residual connections surround both sublayers.

Figure 5 illustrates the shared model architecture and the internal structures of the grouped-query attention and SwiGLU modules.

A.2 Model Configurations

We construct six model configurations by varying the residual-stream width N from 512 to 3072 while fixing the number of Transformer blocks to $L = 32$. The attention head dimension is fixed at 128. Accordingly, the number of query heads increases proportionally with N. We use a 4:1 ratio between query heads and key–value heads, and set the SwiGLU intermediate dimension to $d_{\text{ffn}} = 3.5N$.

Table 1 summarizes the resulting architectural configurations and total parameter counts.

Model	L	N	Q heads	KV heads	d_{ffn}	Params.
GPT-W512	32	512	4	1	1792	0.1524B
GPT-W1024	32	1024	8	2	3584	0.5229B
GPT-W1536	32	1536	12	3	5376	1.1114B
GPT-W2048	32	2048	16	4	7168	1.7450B
GPT-W2560	32	2560	20	5	8960	2.7265B
GPT-W3072	32	3072	24	6	10752	3.9261B

Table 1: Architectural configurations used in the width-scaling experiments. All models contain 32 Transformer blocks, use an attention-head dimension of 128, and maintain a 4:1 query-to-key/value head ratio.

The depth, tokenizer, vocabulary, context length, training-token budget, data mixture, and optimization protocol are held fixed across all configurations. The residual-stream width, number of attention heads, number of key–value heads, feed-forward dimension, and total parameter count therefore scale jointly with N. This construction isolates width-dependent architectural scaling while controlling the training protocol and data budget.

A.3 Pretraining Data

All models are pretrained on the same multi-source text mixture. The mixture contains general web text, educational web documents, deduplicated multi-domain text, WebText-style documents, encyclopedic content, technical question–answer data, and scientific articles. The approximate sampling proportions are provided in Table 2.

Data source	Proportion
C4	35%
FineWeb-Edu	30%
SlimPajama	20%
OpenWebText2	10%
Wikipedia, StackExchange, and arXiv	5%
Total	100%

Table 2: Pretraining data mixture shared by all model configurations. Percentages denote approximate token-level sampling proportions.

A.4 Pretraining Protocol

All model configurations are pretrained from scratch under the same optimization protocol. We use AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$, numerical stability constant $\epsilon = 10^{-8}$, and weight decay 0.1. Weight decay is excluded from embedding parameters, RMSNorm parameters, and bias terms. The learning rate is linearly warmed up for 2,000 optimization steps and subsequently decayed according to a cosine schedule. Gradients are clipped to a maximum norm of 1.0, and training is performed in BF16 precision without dropout.

All models use a context length of 2048 tokens and are trained on a budget of 10 billion tokens. The 10B-token runs serve as the primary controlled experiments in this work; larger-token training budgets are left for subsequent scaling experiments.

Hyperparameter	Setting
Architecture	
Architecture	Decoder-only Transformer
Normalization	RMSNorm, pre-normalization
Position encoding	RoPE
Attention	Grouped-query attention
Feed-forward network	SwiGLU
Tokenizer	Llama 3 tokenizer
Initialization	Normal, $\sigma = 0.02$
Optimizer	AdamW
Adam (β_1, β_2)	$(0.9, 0.95)$
Adam ϵ	10^{-8}
Weight decay	0.1
No weight decay	Embeddings, RMSNorm, biases
Learning-rate schedule	Linear warmup + cosine decay
Warmup	2,000 steps
Peak learning rate	3×10^{-4}
Minimum learning rate	1.5×10^{-7}
Gradient clipping	1.0
Precision	BF16
Dropout	0
Global batch size	480 sequences
Sequence length	2048 tokens
Training budget	10B tokens
Independent runs	8 per model configuration
Training seeds	1234, 520, 314, 315, 3, 1, 5, 31

Table 3: Shared pretraining hyperparameters.

A.5 Independent Training Runs

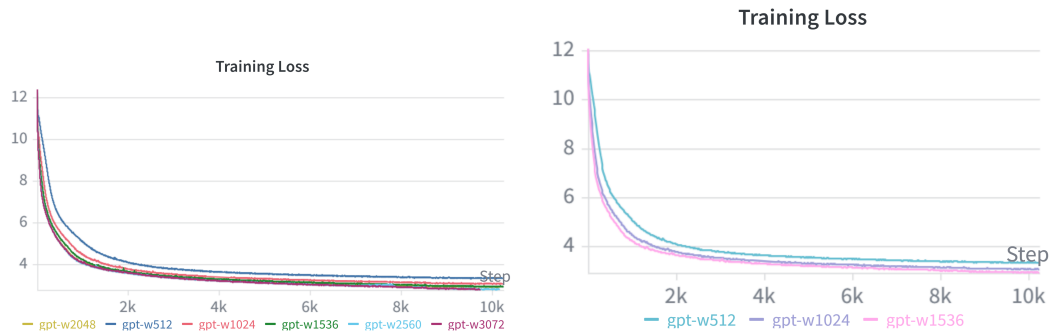
For each model configuration, we perform eight independent pretraining runs using random seeds. Each run starts from an independently initialized parameter state. When controlled by the global random seed, the training-data order and other stochastic training operations are also independently randomized. This produces $6 \times 8 = 48$ independently trained models in total.

Unless otherwise stated, reported results are averaged over the eight independent training runs. Per-run results are retained to quantify the variability induced by parameter initialization and stochastic optimization.

B Training Dynamics and Optimization Stability

B.1 Training Loss

Training loss is computed from the token-level cross-entropy of each optimization batch. Figure 6a shows the complete training-loss trajectories for all six model widths under the shared token budget and optimization schedule. Figure 6b provides a focused view of the three smaller configurations, making their relative convergence behavior easier to distinguish.



(a) All six model widths.

(b) The three smaller widths.

Figure 6: Training-loss trajectories under the shared token budget and optimization schedule. Panel (a) compares all six widths; panel (b) enlarges the three smaller configurations.

Across widths, the curves provide a direct view of optimization progress under the controlled training protocol. Short-timescale fluctuations are expected from minibatch sampling and do not by themselves indicate instability; the relevant diagnostic is the longer-term convergence trend.

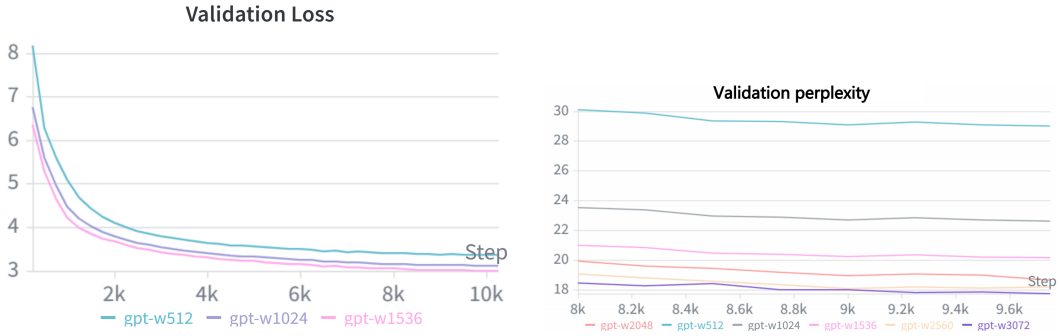
B.2 Validation Loss and Late-Training Perplexity

Validation loss is evaluated on the fixed held-out split at regular intervals using the same tokenization, causal masking, and token-level normalization as in training. Figure 7a shows the validation loss trajectories for the three smaller model configurations.

To summarize validation behavior near the end of training, we compute perplexity as

$$\mathrm{PPL}_{\mathrm{val}} = \exp(\mathcal{L}_{\mathrm{val}}), \tag{2}$$

where $\mathcal{L}_{\mathrm{val}}$ is the mean token-level negative log-likelihood. Figure 7b reports validation perplexity over the late-training checkpoints for all six widths. Because perplexity is an exponential transformation of validation loss, small loss differences can appear more pronounced on the perplexity scale; all comparisons therefore use an identical evaluation corpus and loss normalization.



(a) Validation loss for the three smaller widths.

(b) Late-training perplexity for all six widths.

Figure 7: Held-out performance during pretraining. The two panels use the same validation corpus and token-level loss normalization.

B.3 Gradient Norm and Training Stability

We track the global ℓ_2 norm of the gradients before clipping throughout pretraining. For parameters θ and step-t gradient g_t, the reported quantity is

$$\|g_t\|_2 = \left(\sum_{p \in \theta} \|g_{t,p}\|_2^2 \right)^{1/2}. \tag{3}$$

The norm is measured before applying the maximum-norm threshold of 1.0 so that the diagnostic retains information about gradient spikes. Values above the threshold indicate steps on which clipping rescales the update; they do not indicate that the post-clipping update norm exceeds the threshold.

Figure 8 shows the gradient-norm trajectories for all model widths. We assess stability using the overall scale and trend of the norms, the frequency and duration of excursions above the clipping threshold, and agreement across seeds. Isolated spikes can arise from unusually difficult minibatches, while persistent growth, repeated extreme excursions, non-finite values, or a simultaneous increase in validation loss would be evidence of optimization instability. Reporting the pre-clipping norm together with the loss curves makes it possible to distinguish benign clipped spikes from instability that materially affects convergence.

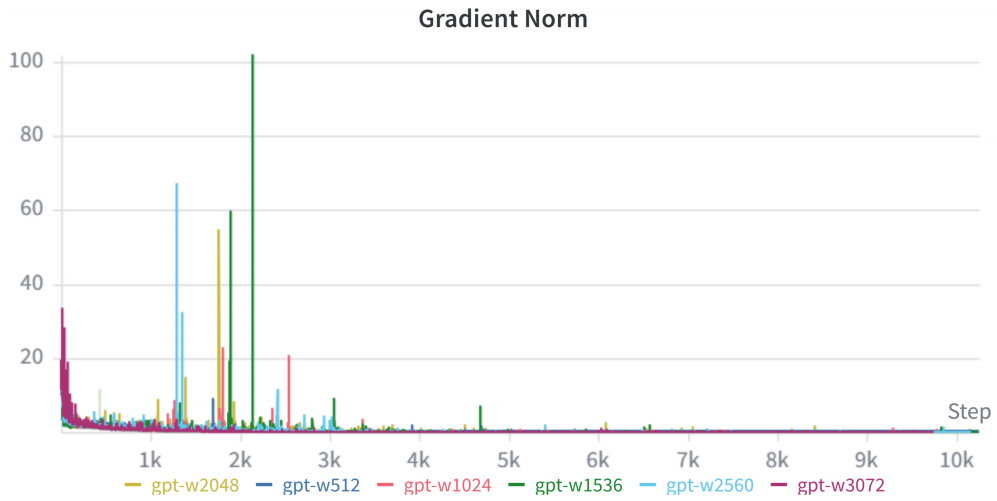


Figure 8: Global gradient ℓ_2 norm before clipping across optimization steps. Curves show means over eight independent runs, shaded regions show one standard deviation across seeds, and the horizontal dashed line marks the clipping threshold of 1.0.

B.4 Pythia Model Configurations

We evaluate five models from the Pythia family. The small-scale regime contains Pythia-14M, Pythia-31M, and Pythia-70M, all of which use six Transformer layers and hidden widths $N \in \{128, 256, 512\}$. The large-scale regime contains Pythia-2.8B and Pythia-6.9B, both of which use 32 Transformer layers and hidden widths $N \in \{2560, 4096\}$. This grouping supports controlled within-depth comparisons of activation-gradient geometry as hidden width increases. The architectural configurations are summarized in Table 4.

Model	Layers	Width N	Heads H	d_{head}	d_{ff}
Pythia-14M	6	128	4	32	512
Pythia-31M	6	256	8	32	1024
Pythia-70M	6	512	8	64	2048
Pythia-2.8B	32	2560	32	80	10240
Pythia-6.9B	32	4096	32	128	16384

Table 4: Architectural configurations of the Pythia models used in our experiments. All configurations use a context length of 2048, a vocabulary of 50,304, and parallel attention/feed-forward residual branches. Here $d_{\text{head}} = N/H$.

All evaluated models use the GPT-NeoX decoder-only architecture with GELU activations, parallel attention and feed-forward residual branches, untied input and output embeddings, and rotary positional embeddings. Rotary embeddings are applied to 25% of each attention head using a rotary

base of $10{,}000$. As shown in Table 4, the feed-forward dimension satisfies $d_{\text{ff}} = 4N$ for all evaluated configurations, whereas the number and dimension of attention heads follow the original Pythia architectures.

For Pythia-14M, Pythia-31M, and Pythia-70M, we evaluate ten independently trained runs for each configuration. These runs share the same architecture and training protocol but differ in their training random seeds. For Pythia-2.8B and Pythia-6.9B, one pretrained checkpoint is evaluated at each width, and variability in the activation-gradient measurements is estimated across independently sampled evaluation subsets.

We additionally evaluate the conventional language-modeling performance of all five Pythia configurations on WikiText-2. All perplexity values are recomputed using the same preprocessing and evaluation protocol. The results are reported in Table 5.

Model	WikiText-2 PPL ↓
Pythia-14M	169.26
Pythia-31M	114.24
Pythia-70M	93.65
Pythia-2.8B	17.71
Pythia-6.9B	15.85

Table 5: Recomputed WikiText-2 perplexity of the evaluated Pythia model configurations. Lower values indicate better language-modeling performance.

As shown in Table 5, WikiText-2 perplexity decreases consistently across the evaluated Pythia configurations as model width and overall capacity increase. These conventional language-modeling results verify that the evaluated models exhibit the expected predictive-performance ordering and complement the activation-gradient geometry analyzed in the main experiments.

B.5 ResNet-20 Experimental Configuration and Accuracy

We evaluate width-dependent behavior on CIFAR-10 and CIFAR-100 using a standard ResNet-20 architecture composed of BasicBlocks. Each input image has resolution $3 \times 32 \times 32$, and the network contains three residual stages with three BasicBlocks per stage. For width multiplier $w \in \{0.5, 1, 2, 4\}$, the channel dimensions of the three stages are $[16w, 32w, 64w]$. The first blocks of the second and third stages perform spatial downsampling using stride 2. Following the original CIFAR ResNet design, we use the parameter-free Option-A shortcut, which applies strided subsampling and zero padding when the spatial resolution and number of channels change.

The complete architectural and optimization settings are summarized in Table 6. Models are trained using standard cross-entropy loss without label smoothing. All convolutional layers are initialized using Kaiming-normal initialization. For every dataset–width configuration, we independently train eight models using random seeds 0 through 7.

Table 7 reports the final test accuracy obtained after $64{,}000$ training steps. For each dataset–width configuration, the final accuracy is summarized as the mean and standard deviation across eight independently trained models. The parameter count includes all trainable parameters, including the dataset-dependent classification head.

As shown in Table 7, increasing the channel-width multiplier consistently improves mean final test accuracy on both CIFAR-10 and CIFAR-100, confirming the expected width-dependent predictive-performance improvement across the evaluated checkpoints.

Setting	Value
Datasets	CIFAR-10, CIFAR-100
Input resolution	$3 \times 32 \times 32$
Architecture	ResNet-20 with BasicBlocks
Blocks per stage	$3, 3, 3$
Width multipliers	$0.5\times, 1\times, 2\times, 4\times$
Stage channels	$[16w, 32w, 64w]$
Downsampling	Stage 2 and 3, stride 2
Shortcut	Option A with zero padding
Loss	Cross-entropy
Label smoothing	None
Initialization	Kaiming normal
Optimizer	SGD
Initial learning rate	0.1
Momentum	0.9
Weight decay	10^{-4}
Batch size	128
Training steps	$64{,}000$
LR scheduler	MultiStepLR
LR milestones	$32{,}000,\ 48{,}000$ steps
LR decay factor	0.1
Learning-rate schedule	$0.1 \to 0.01 \to 0.001$
Evaluation interval	Every $8{,}000$ steps
Logging interval	Every 500 steps
Training seeds	$0, 1, 2, 3, 4, 5, 6, 7$

Table 6: Architecture and training configuration used for the ResNet-20 experiments on CIFAR-10 and CIFAR-100.

Dataset	Width	Parameters	Final test accuracy (%)
CIFAR-10	$0.5\times$	$68{,}050$	88.21 ± 0.21
CIFAR-10	$1\times$	$269{,}722$	91.56 ± 0.30
CIFAR-10	$2\times$	$1{,}073{,}962$	93.61 ± 0.15
CIFAR-10	$4\times$	$4{,}286{,}026$	94.49 ± 0.35
CIFAR-100	$0.5\times$	$71{,}020$	58.25 ± 0.57
CIFAR-100	$1\times$	$275{,}572$	67.19 ± 0.25
CIFAR-100	$2\times$	$1{,}085{,}572$	72.00 ± 0.27
CIFAR-100	$4\times$	$4{,}309{,}156$	75.67 ± 0.27

Table 7: Final test accuracy of the evaluated ResNet-20 width configurations. Results are reported as mean $\pm$ standard deviation across eight independently trained models.

C Computation of Activation-Gradient Alignment Statistics

For each evaluation unit (x_i, y_i), we extract the activation gradient

$$q_i = \nabla_{z_i} \ell(f_{\text{top}}(z_i), y_i)\,, \qquad z_i = f_{\text{bot}}(x_i).$$

Given n activation-gradient vectors, let

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} q_i, \qquad E_{i,:} = (q_i - \widehat{\mu})^\top.$$

The empirical quantities are computed as

$$\widehat{v}_{\|} = \frac{\|E\widehat{\mu}\|_2^2}{n-1}, \qquad \widehat{v}_2 = \frac{\|EE^\top\|_F^2}{(n-1)^2}.$$

Algorithm 1 Estimation of activation-gradient alignment statistics. The Gram-matrix computation avoids explicitly forming the $N \times N$ covariance matrix.

Require: Model $f = f_{\mathrm{top}} \circ f_{\mathrm{bot}}$, evaluation samples $\{(x_i, y_i)\}_{i=1}^{n}$
Ensure: $\widehat{r}_q, \widehat{d}_{\parallel}, \widehat{d}_2, \widehat{d}_{\mathrm{align}}$
for $i = 1, \ldots, n$ **do**
  $z_i \leftarrow f_{\mathrm{bot}}(x_i)$
  $q_i \leftarrow \nabla_{z_i} \ell(f_{\mathrm{top}}(z_i), y_i)$
end for
$\widehat{r}_q \leftarrow n^{-1} \sum_{i=1}^{n} \|q_i\|_2$
$\widehat{\mu} \leftarrow n^{-1} \sum_{i=1}^{n} q_i$
$E_{i,:} \leftarrow (q_i - \widehat{\mu})^{\top}$
$s \leftarrow \|\widehat{\mu}\|_2^2$
$v_{\parallel} \leftarrow \|E\widehat{\mu}\|_2^2/(n-1)$
$v_2 \leftarrow \|EE^{\top}\|_F^2/(n-1)^2$
$\widehat{d}_{\parallel} \leftarrow s^2/v_{\parallel}$
$\widehat{d}_2 \leftarrow s^2/v_2$
$\widehat{d}_{\mathrm{align}} \leftarrow \min\{\widehat{d}_{\parallel}, \widehat{d}_2\}$
return $\widehat{r}_q, \widehat{d}_{\parallel}, \widehat{d}_2, \widehat{d}_{\mathrm{align}}$

Thus,

$$\widehat{d}_{\parallel} = \frac{\|\widehat{\mu}\|_2^4}{\widehat{v}_{\parallel}}, \qquad \widehat{d}_2 = \frac{\|\widehat{\mu}\|_2^4}{\widehat{v}_2}, \qquad \widehat{d}_{\mathrm{align}} = \min\{\widehat{d}_{\parallel}, \widehat{d}_2\}.$$

The identity

$$\mathrm{tr}(\widehat{\Sigma}^2) = \frac{\|EE^{\top}\|_F^2}{(n-1)^2}$$

allows the computation to use an $n \times n$ Gram matrix rather than an $N \times N$ covariance matrix. In implementation, denominators smaller than a numerical tolerance are treated as zero, and the corresponding effective dimension is set to $+\infty$.

D Monte Carlo Estimation of Misalignment Probability

The gradient samples used for Monte Carlo misalignment estimation are independently sampled and do not overlap with those used to estimate d_{align}. For each Monte Carlo trial, we independently sample M training-side and K test-side activation gradients and compute

$$\mu_M = \frac{1}{M} \sum_{i=1}^{M} q_i, \qquad g_K = \frac{1}{K} \sum_{j=1}^{K} \widetilde{q}_j.$$

The empirical misalignment probability is

$$\widehat{p}_{M,K} = \frac{1}{R} \sum_{r=1}^{R} \mathbf{1}\left\{ (\mu_M^{(r)})^{\top} g_K^{(r)} \le 0 \right\}.$$

Sampling is performed with replacement from the two fixed, disjoint gradient pools; repeated gradients are therefore allowed within a trial. Conditional on these pools, the estimator targets the empirical misalignment probability induced by their two empirical distributions. We use $R = 500{,}000$ conditionally independent trials for each model configuration and each pair (M, K). The worst-case Monte Carlo standard error is therefore

$$\widehat{\mathrm{SE}}_{M,K} \le \frac{1}{2\sqrt{R}} \approx 7.1 \times 10^{-4}.$$

The reported plug-in certificate estimates are computed after Algorithm 2 as

$$\widehat{\Xi}_{M,K} = \frac{1}{M\widehat{d}_{\parallel}} + \frac{1}{K\widehat{d}_{\parallel}} + \frac{1}{MK\widehat{d}_2},$$

Algorithm 2 Monte Carlo estimation of the activation-gradient misalignment probability. Training-side and test-side gradients are sampled independently with replacement in every trial. Conditional on the two fixed pools, the Monte Carlo trials are independent.

Require: Fixed, disjoint gradient pools $\mathcal{Q}_{\mathrm{tr}}$ and $\mathcal{Q}_{\mathrm{te}}$, sample sizes M, K, trials R
Ensure: $\widehat{p}_{M,K}$ and $\widehat{\mathrm{SE}}_{M,K}$
$C \leftarrow 0$
for $r = 1, \ldots, R$ **do**
  Sample $\{q_i\}_{i=1}^{M}$ with replacement from $\mathcal{Q}_{\mathrm{tr}}$
  Sample $\{\widetilde{q}_j\}_{j=1}^{K}$ with replacement from $\mathcal{Q}_{\mathrm{te}}$
  $\mu_M \leftarrow M^{-1} \sum_{i=1}^{M} q_i$
  $g_K \leftarrow K^{-1} \sum_{j=1}^{K} \widetilde{q}_j$
  $C \leftarrow C + \mathbf{1}\{\mu_M^{\top} g_K \le 0\}$
end for
$\widehat{p}_{M,K} \leftarrow C/R$
$\widehat{\mathrm{SE}}_{M,K} \leftarrow \sqrt{\widehat{p}_{M,K}(1-\widehat{p}_{M,K})/R}$
return $\widehat{p}_{M,K}, \widehat{\mathrm{SE}}_{M,K}$

$$\widehat{B}^{\mathrm{exact}}_{M,K} = \frac{\widehat{\Xi}_{M,K}}{1+\widehat{\Xi}_{M,K}}, \qquad \widehat{B}^{\mathrm{align}}_{M,K} = \frac{1}{1+n_{\mathrm{eff}}\widehat{d}_{\mathrm{align}}},$$

where

$$n_{\mathrm{eff}} = \frac{MK}{M+K+1}.$$

For configurations with multiple training seeds, both procedures are applied separately to each trained model. Estimates are first aggregated across independently sampled evaluation subsets within each model and are subsequently summarized across training seeds.

E Component-wise Decomposition of the Effective Alignment Dimension

Recall that

$$d_{\mathrm{align}} = \min\{d_{\|}, d_2\}.$$

To examine which component controls the effective alignment dimension, we report $d_{\|}$, d_2, and d_{align} for the controlled LLaMA-style models. For Pythia, the available table contains only aggregate component summaries; we therefore do not infer an aggregate minimum or bottleneck frequency from those summaries.

Controlled LLaMA-style models. Table 8 reports the component decomposition for the controlled width-scaling family.

Width N	$d_{\|}$	d_2	d_{align}	d_2 bottleneck
512	11.736 ± 0.694	3.944 ± 0.221	3.944 ± 0.221	100%
1024	10.620 ± 0.749	4.418 ± 0.319	4.418 ± 0.319	100%
1536	11.347 ± 0.748	6.261 ± 0.885	6.261 ± 0.885	100%
2048	9.729 ± 0.746	6.691 ± 0.788	6.691 ± 0.788	100%
2560	9.056 ± 0.432	7.767 ± 0.991	7.635 ± 0.883	80%
3072	11.740 ± 0.748	9.525 ± 0.467	9.523 ± 0.463	95%

Table 8: Component-wise decomposition of the effective alignment dimension across the controlled LLaMA-style model widths. The final column reports the frequency with which d_2 is the active bottleneck.

The covariance-energy component d_2 is the dominant bottleneck over most of the investigated width range. It determines d_{align} consistently up to $N = 2048$, while $d_{\|}$ becomes limiting only occasionally

at the two largest widths. Although $d_{\|}$ is not monotonic, d_2 and the resulting d_{align} exhibit a clear overall increase with width, with d_{align} increasing from 3.944 at $N = 512$ to 9.523 at $N = 3072$.

Pythia models. Table 9 reports the corresponding aggregate component summaries within each fixed-depth group.

Width N	$d_{\|}$	d_2
6-layer Pythia models		
128	5.5915	3.4390
256	7.6445	14.2963
512	19.5501	14.3800
32-layer Pythia models		
2560	7.9527	11.0866
4096	9.3290	10.0440

Table 9: Aggregate component summaries for the Pythia models. Comparisons are made within each fixed-depth group. The table does not derive an aggregate d_{align} or a bottleneck frequency from the two component means.

For Pythia, the valid aggregation order is to compute

$$d_{\text{align}}^{(r)} = \min\left\{ d_{\|}^{(r)}, d_2^{(r)} \right\}$$

for each evaluation unit r and only then aggregate the resulting $d_{\text{align}}^{(r)}$ values. Bottleneck frequencies must likewise be computed from the same evaluation units. Neither quantity can in general be recovered from the two aggregate component means in Table 9; accordingly, no Pythia bottleneck claim is made from this table.

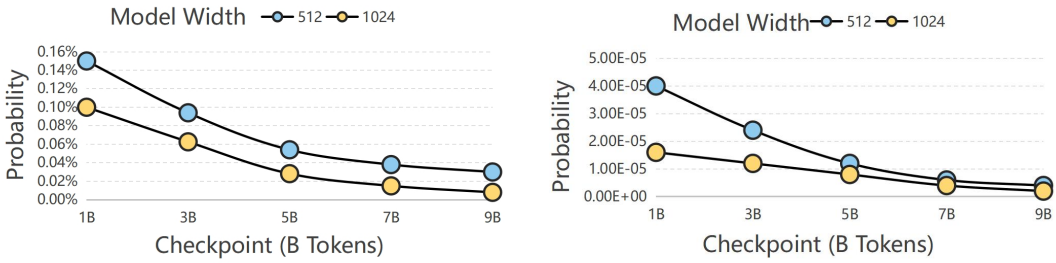
F Evolution of Gradient-Misalignment Probability during Pretraining

We next examine how finite-sample gradient misalignment evolves during pretraining and whether the width-dependent alignment advantage persists throughout the training trajectory. For a model of width N evaluated after T pretraining tokens, we estimate

$$\widehat{p}_{M,K}(N;T) = \frac{1}{R}\sum_{r=1}^{R} \mathbf{1}\left\{ \left(\mu_{M,N,T}^{(r)}\right)^{\top} g_{K,N,T}^{(r)} \le 0 \right\}, \tag{4}$$

where $\mu_{M,N,T}^{(r)}$ and $g_{K,N,T}^{(r)}$ are independently sampled training- and test-averaged gradient signals in the r-th trial. Each sampled sequence contains 2048 tokens, and the evaluation protocol is held fixed across model widths and checkpoints within each comparison.

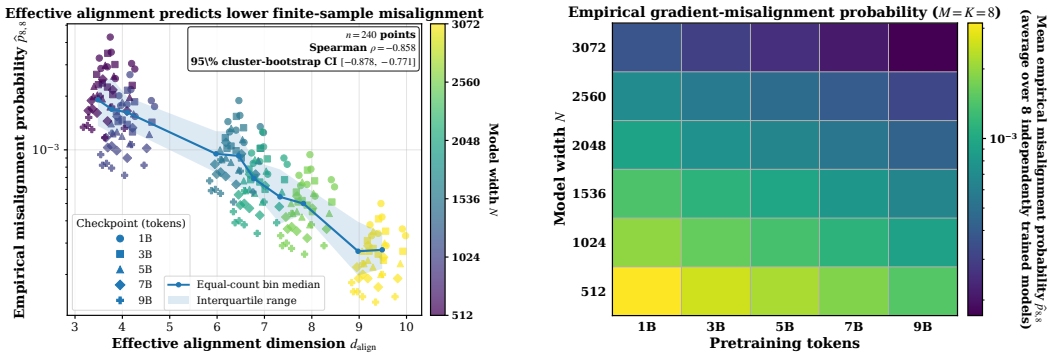
We first evaluate the width-512 and width-1024 LLaMA-style models at checkpoints corresponding to 1, 3, 5, 7, and 9 billion pretraining tokens. We consider two finite-sample budgets, $M = K = 8$ and $M = K = 16$, in order to determine whether the checkpoint-wise behavior persists under different levels of gradient-estimation uncertainty.



(a) $M = K = 8$.

(b) $M = K = 16$.

Figure 9: Evolution of the empirical finite-sample gradient-misalignment probability during pretraining. Width-512 and width-1024 LLaMA-style models are evaluated at checkpoints corresponding to 1, 3, 5, 7, and 9 billion pretraining tokens. Each sampled sequence contains 2048 tokens. Panel (a) uses $M = K = 8$, whereas panel (b) uses $M = K = 16$. For both sampling budgets, the empirical misalignment probability decreases with pretraining progress and remains lower for the wider model.



(a) Relationship between d_{align} and $\widehat{p}_{8,8}$. Color denotes model width and marker shape denotes the pretraining checkpoint.

(b) Empirical misalignment over the complete width–checkpoint grid. Each cell reports the mean over eight independently trained models.

Figure 10: Extended analysis of finite-sample gradient misalignment for $M = K = 8$. Panel (a) retains all 240 width–checkpoint–seed observations. The solid curve reports medians over equal-count bins, and the shaded region denotes the corresponding interquartile range. Panel (b) reports the complete 6×5 width–checkpoint grid. Each of the 30 cells is computed by first estimating $\widehat{p}_{8,8}$ separately for each trained model and then averaging across eight training seeds. Higher d_{align} is strongly associated with lower empirical misalignment (Spearman $\rho = -0.858$; trajectory-clustered bootstrap 95% CI $[-0.878, -0.771]$).

As shown in Figure 9, the empirical gradient-misalignment probability decreases consistently as pretraining progresses. For $M = K = 8$, the probability for the width-512 model decreases from approximately 0.15% at the 1B-token checkpoint to approximately 0.03% at the 9B-token checkpoint. The width-1024 model follows the same decreasing trajectory but remains below the narrower model at every checkpoint, decreasing from approximately 0.10% to below 0.01%.

The same qualitative behavior is observed for $M = K = 16$, but at a substantially smaller absolute probability scale. Increasing the sampling budget reduces the uncertainty of both the training- and test-averaged gradient estimates, while the width-1024 model continues to exhibit a lower misalignment probability throughout pretraining. The width-dependent advantage therefore remains visible even when the misalignment event becomes rare.

Because the two widths are compared at matched pretraining-token checkpoints and under identical sampling budgets within each panel, the observed difference cannot be attributed to additional pretraining data or a larger evaluation sample for the wider model. Instead, the results indicate that

the width-dependent alignment advantage emerges early and persists throughout the investigated training trajectory.

Complete width–checkpoint analysis. We further test whether this behavior extends beyond the two representative widths in Figure 9. For $M = K = 8$, we evaluate all six LLaMA-style widths

$$N \in \{512, 1024, 1536, 2048, 2560, 3072\} \tag{5}$$

at the same five pretraining checkpoints. Each width is represented by eight independently trained models, resulting in

$$6 \text{ widths} \times 5 \text{ checkpoints} \times 8 \text{ training seeds} = 240 \tag{6}$$

model–checkpoint observations.

For the heatmap, we first compute $\widehat{p}_{8,8}$ separately for each trained model and then average the eight model-level estimates within each width–checkpoint pair. The resulting grid therefore contains $6 \times 5 = 30$ cells. The representative trajectories above and the complete grid are estimated from independently sampled evaluation subsets, so their absolute values need not coincide exactly; the relevant comparison is the width- and checkpoint-wise ordering within each analysis.

Figure 10(b) shows a consistent two-dimensional pattern. At every fixed checkpoint, increasing model width reduces the empirical misalignment probability. At every fixed width, additional pretraining also reduces the probability. In the complete-grid evaluation, the mean probability for width 512 decreases from 3.16×10^{-3} at 1B tokens to 1.47×10^{-3} at 9B tokens, whereas the corresponding probability for width 3072 decreases from 3.69×10^{-4} to 1.71×10^{-4}.

Consequently, the width-3072 models exhibit an approximately 8.6-fold smaller misalignment probability than the width-512 models at both the 1B- and 9B-token checkpoints. The ordering is not restricted to the two extreme widths: each intermediate width produces a trajectory lying between those of its narrower and wider counterparts.

The complete grid also shows that width and pretraining progress play complementary roles. Additional pretraining shifts the trajectory of every width toward lower misalignment probability, whereas increasing width shifts the complete checkpoint trajectory downward. In particular, sufficiently wide models at early checkpoints can exhibit lower misalignment than much narrower models after substantially more pretraining. The observed width-dependent improvement therefore cannot be reduced to training progress alone.

Effective alignment dimension predicts misalignment. The theoretical analysis identifies the gradient mean–covariance geometry, rather than nominal width itself, as the quantity directly governing finite-sample directional transfer. We therefore retain every width–checkpoint–seed configuration as an individual observation and compare

$$d_{\text{align}}(N; T, s) \qquad \text{with} \qquad \widehat{p}_{8,8}(N; T, s), \tag{7}$$

where s denotes the training seed. Unlike the heatmap, this analysis does not average over training seeds and therefore retains all 240 observations.

As shown in Figure 10(a), larger d_{align} is strongly associated with lower empirical misalignment. Across all widths, checkpoints, and training seeds, the Spearman rank correlation is

$$\rho_{\text{S}} = -0.858. \tag{8}$$

The five checkpoints obtained from the same trained model are statistically dependent. We therefore quantify uncertainty by resampling complete model trajectories indexed by (N, s) rather than treating all 240 points as independent. A trajectory-clustered bootstrap with $5{,}000$ replicates gives the 95% confidence interval

$$[-0.878, -0.771]. \tag{9}$$

Importantly, this association is not produced solely by pooling different pretraining stages. When the Spearman correlation is computed separately at each checkpoint, it remains strongly negative:

$$\rho_{\text{S}} = -0.959,\ -0.953,\ -0.952,\ -0.948,\ -0.944 \tag{10}$$

at 1, 3, 5, 7, and 9 billion tokens, respectively. Thus, even after pretraining progress is held fixed, models with larger effective alignment dimension consistently exhibit lower finite-sample misalignment.

These results distinguish the proposed mechanism from a direct dependence on parameter width alone. Model width and pretraining progress alter the underlying gradient mean and covariance, while d_{align} summarizes how favorable the resulting geometry is for finite-sample train–test directional transfer. The empirical evidence therefore supports the mechanism

$$(N, T) \longrightarrow d_{\mathrm{align}}(N; T) \longrightarrow \widehat{p}_{M,K}(N; T), \tag{11}$$

rather than merely showing two unrelated correlations with model scale.

Pretraining progress and model width consequently have complementary effects. For a fixed model width and sampling budget, additional pretraining reduces the probability of estimating a direction that is misaligned with an independent test estimate. For a fixed checkpoint and sampling budget, increasing width shifts the model toward a lower-misalignment regime. Over the investigated model family, the results support

$$T \uparrow \implies \widehat{p}_{M,K}(N; T) \downarrow, \qquad N \uparrow \implies \widehat{p}_{M,K}(N; T) \downarrow, \tag{12}$$

for both $(M, K) = (8, 8)$ and $(M, K) = (16, 16)$ in the representative-width comparison, and across the complete six-width grid for $M = K = 8$.

More directly, the model-level analysis supports

$$d_{\mathrm{align}}(N; T, s) \uparrow \implies \widehat{p}_{8,8}(N; T, s) \downarrow . \tag{13}$$

This is precisely the qualitative relationship predicted by the finite-sample alignment certificate: models with more favorable gradient mean–covariance geometry are less likely to produce disagreeing training and test directions from finite samples.

G Problem Setup and Notation

Core Notation

$\mathcal{D}$	Population distribution over input–label pairs, $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$.
M	Number of independently sampled training examples in $S_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{M} \sim \mathcal{D}^M$.
K	Number of independently sampled test examples in $S_{\text{test}} = \{(\widetilde{x}_j, \widetilde{y}_j)\}_{j=1}^{K} \sim \mathcal{D}^K$, where S_{test} is independent of S_{train}.
f_{bot}	Lower part of the trained network, $f_{\text{bot}} : \mathcal{X} \to \mathbb{R}^N$.
f_{top}	Upper part of the trained network, $f_{\text{top}} : \mathbb{R}^N \to \mathbb{R}^C$, with $f = f_{\text{top}} \circ f_{\text{bot}}$.
N	Hidden width at the selected residual-module insertion point.
$z(x)$	Hidden representation at the insertion point, $z(x) = f_{\text{bot}}(x) \in \mathbb{R}^N$.
$q(x, y)$	Activation gradient at the insertion point, $q(x, y) = \nabla_z \ell(f_{\text{top}}(z), y)\big\|_{z=z(x)} \in \mathbb{R}^N$.
$\bar{\mu}$	Population mean activation gradient, $\bar{\mu} = \mathbb{E}_{(x,y)\sim\mathcal{D}}[q(x, y)] \in \mathbb{R}^N$, with $\bar{\mu} \neq 0$.
Σ	Activation-gradient covariance, $\Sigma = \mathrm{Cov}(q(x, y)) = \mathbb{E}[(q(x, y) - \bar{\mu})(q(x, y) - \bar{\mu})^\top] \in \mathbb{R}^{N\times N}$.
μ_M	Training-set activation-gradient average, $\mu_M = M^{-1} \sum_{i=1}^{M} q(x_i, y_i)$.
g_K	Independent test-set activation-gradient average, $g_K = K^{-1} \sum_{j=1}^{K} q(\widetilde{x}_j, \widetilde{y}_j)$.
$d_{\parallel}$	Directional effective dimension, $d_{\parallel} = \|\bar{\mu}\|_2^4 / (\bar{\mu}^\top \Sigma \bar{\mu})$.
d_2	Covariance-energy effective dimension, $d_2 = \|\bar{\mu}\|_2^4 / \|\Sigma\|_F^2 = \|\bar{\mu}\|_2^4 / \mathrm{tr}(\Sigma^2)$.
d_{align}	Effective alignment dimension, $d_{\text{align}} = \min\{d_{\parallel}, d_2\}$.
n_{eff}	Effective sample size, $n_{\text{eff}} = MK/(M + K + 1)$.
$\mathcal{E}_{M,K}$	Activation-gradient misalignment event, $\mathcal{E}_{M,K} = \{\mu_M^\top g_K \leq 0\}$.
$\widetilde{f}_{S_{\text{train}},\eta}$	Training-sample-dependent function-preserving residual-expansion path satisfying $\widetilde{f}_{S_{\text{train}},0} = f$.
$u_{S_{\text{train}}}$	First-order residual perturbation induced by the expansion path, $u_{S_{\text{train}}}(z) := \frac{\partial}{\partial \eta} h_{S_{\text{train}},\eta}(z)\big\|_{\eta=0}$.

This section introduces the activation-space formulation used throughout the theoretical analysis. We consider a trained network, a candidate insertion point, and a function-preserving residual expansion. All alignment quantities are defined exclusively from the activation-gradient signal at the selected insertion point.

Data and risks. Let $\mathcal{D}$ be a probability distribution over $\mathcal{X} \times \mathcal{Y}$. Let

$$S_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{M} \sim \mathcal{D}^M$$

and

$$S_{\text{test}} = \{(\widetilde{x}_j, \widetilde{y}_j)\}_{j=1}^{K} \sim \mathcal{D}^K$$

be mutually independent training and test samples, where $M, K \in \mathbb{N}_+$.

Let

$$\ell : \mathbb{R}^C \times \mathcal{Y} \to \mathbb{R}$$

be the loss function. For any predictor $f : \mathcal{X} \to \mathbb{R}^C$, define

$$L_{\text{train}}(f) := \frac{1}{M} \sum_{i=1}^{M} \ell\big(f(x_i), y_i\big),$$

$$L_{\text{test}}(f) := \frac{1}{K} \sum_{j=1}^{K} \ell\big(f(\widetilde{x}_j), \widetilde{y}_j\big),$$

$$R(f) := \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell\big(f(x), y\big)\right].$$

Here, L_{train} and L_{test} are empirical risks evaluated on finite independent samples, whereas R denotes the population risk under $\mathcal{D}$.

Network decomposition. Fix a candidate insertion point in a trained network and decompose the original model as

$$f = f_{\text{top}} \circ f_{\text{bot}},$$

where

$$f_{\text{bot}} : \mathcal{X} \to \mathbb{R}^N, \qquad f_{\text{top}} : \mathbb{R}^N \to \mathbb{R}^C.$$

For an input $x \in \mathcal{X}$, let

$$z(x) := f_{\text{bot}}(x) \in \mathbb{R}^N$$

denote the hidden representation at the selected insertion point. Thus,

$$f(x) = f_{\text{top}}\big(z(x)\big),$$

where N is the hidden width at that insertion point.

Function-preserving residual expansion. Given the training sample S_{train}, consider a differentiable residual-expansion path

$$\eta \longmapsto \widetilde{f}_{S_{\text{train}},\eta}, \qquad \eta \geq 0,$$

of the form

$$\widetilde{f}_{S_{\text{train}},\eta}(x) = f_{\text{top}}\left(z(x) + h_{S_{\text{train}},\eta}\big(z(x)\big)\right),$$

where

$$h_{S_{\text{train}},0}(z) = 0 \qquad \text{for every } z \in \mathbb{R}^N.$$

Consequently,

$$\widetilde{f}_{S_{\text{train}},0}(x) = f_{\text{top}}\big(z(x)\big) = f(x),$$

so the expansion preserves the original input–output function at $\eta = 0$.

The training sample determines the initial residual direction, while the scalar η controls the magnitude of the expansion.

Activation-gradient signal. For a sample $(x, y) \sim \mathcal{D}$, define the activation gradient at the insertion point by

$$q(x, y) := \nabla_z \ell\left(f_{\text{top}}(z), y\right)\Big|_{z=z(x)} \in \mathbb{R}^N.$$

The random vector $q(x, y)$ is the loss-gradient signal in the hidden representation space at the selected insertion point.

Throughout the paper, all population moments, finite-sample averages, alignment events, and effective dimensions are defined from $q(x, y)$.

Population activation-gradient moments. Define the population mean activation gradient by

$$\bar{\mu} := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[q(x, y)\right] \in \mathbb{R}^N,$$

and its covariance by

$$\Sigma := \text{Cov}\big(q(x, y)\big) = \mathbb{E}\left[\big(q(x, y) - \bar{\mu}\big)\big(q(x, y) - \bar{\mu}\big)^\top\right] \in \mathbb{R}^{N\times N}.$$

The finite-sample alignment analysis assumes

$$\mathbb{E}\left[\|q(x, y)\|_2^2\right] < \infty \qquad \text{and} \qquad \bar{\mu} \neq 0.$$

The finite-second-moment condition ensures that both $\bar{\mu}$ and Σ are well defined.

Training and test activation-gradient averages. Define the training-set activation-gradient average by

$$\mu_M := \frac{1}{M}\sum_{i=1}^{M} q(x_i, y_i),$$

and the independently evaluated test-set activation-gradient average by

$$g_K := \frac{1}{K}\sum_{j=1}^{K} q(\widetilde{x}_j, \widetilde{y}_j).$$

Both quantities are unbiased estimators of the same population activation-gradient signal:

$$\mathbb{E}[\mu_M] = \mathbb{E}[g_K] = \bar{\mu}.$$

Moreover, μ_M and g_K are independent because they are computed from mutually independent samples.

Activation-gradient alignment event. The inner product

$$\mu_M^\top g_K$$

measures the directional agreement between the activation gradients estimated from the training and test samples.

The event

$$\mu_M^\top g_K > 0$$

corresponds to positive train–test activation-gradient alignment. The conservative failure event is

$$\mathcal{E}_{M,K} := \left\{\mu_M^\top g_K \le 0\right\}.$$

Connection to residual-expansion descent. For the training-sample-dependent residual-expansion path

$$\widetilde{f}_{S_{\text{train}},\eta}(x) = f_{\text{top}}\left(z(x) + h_{S_{\text{train}},\eta}(z(x))\right),$$

define its first-order residual perturbation by

$$u_{S_{\text{train}}}(z) := \left.\frac{\partial}{\partial\eta} h_{S_{\text{train}},\eta}(z)\right|_{\eta=0}.$$

Because $h_{S_{\text{train}},0}(z) = 0$, the chain rule gives

$$\left.\frac{\mathrm{d}}{\mathrm{d}\eta} L_{\text{test}}\big(\widetilde{f}_{S_{\text{train}},\eta}\big)\right|_{\eta=0} = \frac{1}{K}\sum_{j=1}^{K} \widetilde{q}_j^\top\, u_{S_{\text{train}}}(\widetilde{z}_j), \tag{14}$$

where

$$\widetilde{z}_j = f_{\text{bot}}(\widetilde{x}_j), \qquad \widetilde{q}_j = q(\widetilde{x}_j, \widetilde{y}_j).$$

Equation equation 14 is the general first-order derivative for a differentiable function-preserving residual-expansion path. It does not, by itself, equal $-\mu_M^\top g_K$.

Residual-realizability condition. Following the direct train–test residual-expansion framework, we assume that the training-selected residual perturbation satisfies

$$\frac{1}{K}\sum_{j=1}^{K} \widetilde{q}_j^\top\, u_{S_{\text{train}}}(\widetilde{z}_j) = -\mu_M^\top g_K. \tag{15}$$

A sufficient realization is

$$u_{S_{\text{train}}}(z) \equiv -\mu_M,$$

whenever the residual tangent family contains this constant activation-space perturbation. Indeed, in this case

$$\frac{1}{K}\sum_{j=1}^{K} \widetilde{q}_j^\top(-\mu_M) = -\mu_M^\top g_K.$$

Under this condition,

$$\frac{\mathrm{d}}{\mathrm{d}\eta} L_{\mathrm{test}}\big(\widetilde{f}_{S_{\mathrm{train}},\eta}\big)\Big|_{\eta=0} = -\mu_M^\top g_K. \tag{16}$$

Therefore,

$$\mu_M^\top g_K > 0 \quad \Longrightarrow \quad \frac{\mathrm{d}}{\mathrm{d}\eta} L_{\mathrm{test}}\big(\widetilde{f}_{S_{\mathrm{train}},\eta}\big)\Big|_{\eta=0} < 0.$$

Hence, positive train–test activation-gradient alignment implies that the training-selected residual perturbation is a strict first-order descent direction for the independently evaluated test risk.

More precisely, differentiability at $\eta = 0$ gives

$$L_{\mathrm{test}}\big(\widetilde{f}_{S_{\mathrm{train}},\eta}\big) = L_{\mathrm{test}}(f) - \eta \mu_M^\top g_K + o(\eta). \tag{17}$$

Equivalently, defining

$$\Delta_R^{\mathrm{test}}(\eta) := L_{\mathrm{test}}(f) - L_{\mathrm{test}}\big(\widetilde{f}_{S_{\mathrm{train}},\eta}\big),$$

we obtain

$$\Delta_R^{\mathrm{test}}(\eta) = \eta \mu_M^\top g_K + o(\eta).$$

Consequently, if $\mu_M^\top g_K > 0$, then there exists $\eta_0 > 0$ such that

$$\Delta_R^{\mathrm{test}}(\eta) > 0 \qquad \text{for every } 0 < \eta \le \eta_0.$$

Effective alignment dimensions. Define the directional effective dimension by

$$d_{\|} := \frac{\|\bar{\mu}\|_2^4}{\bar{\mu}^\top \Sigma \bar{\mu}},$$

and the covariance-energy effective dimension by

$$d_2 := \frac{\|\bar{\mu}\|_2^4}{\|\Sigma\|_F^2} = \frac{\|\bar{\mu}\|_2^4}{\mathrm{tr}(\Sigma^2)}.$$

Whenever a denominator is zero, the corresponding effective dimension is understood to be $+\infty$.

The quantity $d_{\|}$ measures the squared population activation-gradient signal relative to covariance in the population signal direction. The quantity d_2 measures the same signal relative to the total covariance energy.

Define the effective alignment dimension by

$$d_{\mathrm{align}} := \min\left\{d_{\|}, d_2\right\}.$$

This provides a conservative scalar summary of the two relative-noise contributions governing finite-sample activation-gradient alignment.

Effective sample size. Define

$$n_{\mathrm{eff}} := \frac{MK}{M + K + 1}.$$

This quantity combines the training and test sample sizes into the finite-sample factor appearing in the alignment guarantee. It is symmetric in M and K and increases with either sample size.

In the matched-budget setting $M = K = n$,

$$n_{\mathrm{eff}} = \frac{n^2}{2n + 1}.$$

H Residual Expansion Setup and Connection to Prior Work

We spell out the residual insertion and the direct train–test route inherited from Cheng et al. (2026). This makes explicit which ingredients are supplied by the prior work and which alignment step is replaced here.

A concrete insertion and the two comparison models. Fix the trained reference model $f^*_{\mathrm{old}} = f_{\mathrm{top}} \circ f_{\mathrm{bot}}$ and an admissible insertion point independently of the training and test samples used below, and write $z(x) = f_{\mathrm{bot}}(x) \in \mathbb{R}^N$. A residual insertion retains the identity path and changes the hidden state from z to $z + h_\theta(z)$, giving $f_\theta(x) = f_{\mathrm{top}}(z(x) + h_\theta(z(x)))$. For example, let $h_{U,V,b}(z) = V\psi_U(z) + b$, with $V = b = 0$ at initialization. Then $h_{U,0,0} \equiv 0$, so the enlarged network computes exactly f^*_{old} at initialization. Here "residual" refers to the additive skip structure $z + h_\theta(z)$; a replacement map $f_{\mathrm{top}}(h_\theta(z))$ has no such identity path and is not the insertion considered here.

For the activation gradient $q(x, y) = \nabla_z \ell(f_{\mathrm{top}}(z), y)|_{z=z(x)}$ defined above, let μ_M and g_K be its training and independent-test averages over samples of sizes M and K. In this example the training-selected bias path $b(\eta) = -\eta\mu_M, V(\eta) = 0$, induces $f_{S,\eta}(x) = f_{\mathrm{top}}(z(x) - \eta\mu_M)$. The *expansion direction* is the tangent $\partial_\eta h_{S,\eta}(z)|_{\eta=0} = -\mu_M$; it is a direction in the newly inserted residual branch, not a change to the original model before insertion. The chain rule gives

$$\frac{d}{d\eta} L_{\mathrm{train}}(f_{S,\eta})\bigg|_0 = -\|\mu_M\|_2^2, \qquad \frac{d}{d\eta} L_{\mathrm{test}}(f_{S,\eta})\bigg|_0 = -\mu_M^\top g_K. \tag{18}$$

For a sufficiently small positive step η_S, the *empirical jumpboard model* $\widetilde{f}_S := f_{S,\eta_S}$ is an auxiliary training-selected comparison model. The *final expanded model* $f_{\mathrm{new}} \in \mathcal{H}_{\mathrm{new}}$ is obtained after optimizing the expanded network and is required by the selection condition to satisfy $L_{\mathrm{train}}(f_{\mathrm{new}}) \le L_{\mathrm{train}}(\widetilde{f}_S) - \Delta_{\mathrm{ERM}}$, with $\Delta_{\mathrm{ERM}} \ge 0$. The constant tangent above is exactly realizable when the new branch has an additive output parameter (or a constant feature). Otherwise the corresponding parameter-space tangent and its realizability condition must be used; the activation gradient alone does not certify an arbitrary inserted module.

What the earlier generalization result proves. Under the bounded-loss, norm-control, and normalization hypotheses of Cheng et al. (2026), their Lemma 1 bounds the empirical Rademacher complexity of the loss class. Using d_{par} for their parameter-count dimension, to distinguish it from d_{align}, the bound is, for every $\rho > 0$,

$$\widehat{\mathfrak{R}}_S(\ell \circ \mathcal{H}_{\mathrm{new}}) \le B_\ell \sqrt{\frac{2d_{\mathrm{par}}}{m} \log\left(1 + \frac{2\bar{b}LL_\ell\Lambda_{\max}}{\rho}\right)} + \rho. \tag{19}$$

Here m is the sample size, B_ℓ bounds the loss, and $\bar{b}, L, L_\ell, \Lambda_{\max}$ are the parameter-radius, depth, loss-Lipschitz, and layerwise amplification quantities specified in that lemma. Its proof bounds normalized hidden states, propagates layerwise Lipschitz constants, covers the norm-controlled parameter set, and converts the resulting loss-class cover into equation 19. Their Corollary 2 then applies a uniform Rademacher generalization inequality: with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{H}_{\mathrm{new}}} |R(f) - L_S(f)| \le \epsilon_m,$$

$$\epsilon_m := 2B_\ell \sqrt{\frac{2d_{\mathrm{par}}}{m} \log\left(1 + \frac{2\bar{b}LL_\ell\Lambda_{\max}}{\rho}\right)} + 2\rho + 3B_\ell \sqrt{\frac{\log(2/\delta)}{2m}}. \tag{20}$$

Applying this corollary separately with $m = M$ and $m = K$ yields the radii ϵ_M and ϵ_K in Assumption 3 whenever its cited hypotheses hold for the chosen $\mathcal{H}_{\mathrm{new}}$. Thus that assumption can be discharged by the earlier theorem in its applicable model and loss class. Uniformity is essential: it covers both the data-dependent $\widetilde{f}_S$ and f_{new}.

From the jumpboard to test risk. Let $\Delta_R^{\mathrm{test}} := L_{\mathrm{test}}(f^*_{\mathrm{old}}) - L_{\mathrm{test}}(\widetilde{f}_S)$. On the intersection of the two uniform-generalization events, which has probability at least $1 - 2\delta$, one has simultaneously for every $f \in \mathcal{H}_{\mathrm{new}}$ that $|L_{\mathrm{train}}(f) - L_{\mathrm{test}}(f)| \le \epsilon_M + \epsilon_K$. Consequently,

$$\begin{aligned} L_{\mathrm{test}}(f_{\mathrm{new}}) &\le L_{\mathrm{train}}(f_{\mathrm{new}}) + \epsilon_M + \epsilon_K \\ &\le L_{\mathrm{train}}(\widetilde{f}_S) - \Delta_{\mathrm{ERM}} + \epsilon_M + \epsilon_K \\ &\le L_{\mathrm{test}}(f^*_{\mathrm{old}}) - \Delta_R^{\mathrm{test}} - \Delta_{\mathrm{ERM}} + 2(\epsilon_M + \epsilon_K). \end{aligned} \tag{21}$$

This is the risk-comparison mechanism of Cheng et al. (2026, Theorem 3). That theorem states a $1 - 3\delta$ guarantee when the alignment event is also counted; the algebraic comparison above uses

only the two uniform events. The inequality itself does not make Δ_R^{test} positive. Their Corollary 3 supplies the local bridge: for a realizable path with a sufficiently small positive step,

$$\Delta_R^{\text{test}}(\eta) = \eta \mu_M^\top g_K + o(\eta), \tag{22}$$

so $\mu_M^\top g_K > 0$ makes the finite-test jumpboard margin positive locally. The step must be small enough for the realized path; a positive derivative alone does not guarantee improvement at an arbitrary finite step. Finally, their Theorem 5 bounded the failure probability under the covariance-spectrum condition $\lambda_{\max}(\Sigma) \le C_\Sigma \tau^2$:

$$\Pr(\mu_M^\top g_K \le 0) \le \frac{4C_\Sigma \tau^2}{M\|\bar{\mu}\|_2^2} + \frac{4C_\Sigma \tau^2}{K\|\bar{\mu}\|_2^2} + \frac{4C_\Sigma \tau^2 \operatorname{tr}(\Sigma)}{MK\|\bar{\mu}\|_2^4}. \tag{23}$$

Its proof controls separate deviations of the two empirical gradients from $\bar{\mu}$; additional total-variance and signal-growth conditions yield its width-dependent rate. Thus the inherited chain is: uniform generalization, direct train–test comparison, a local alignment-to-margin bridge, and control of alignment failure.

The step replaced in this paper. Our Theorem 1 replaces this last probability-control step. Independence of the two samples gives $\mathbb{E}[\mu_M^\top g_K] = \|\bar{\mu}\|_2^2$ and

$$\operatorname{Var}(\mu_M^\top g_K) = \left(\frac{1}{M} + \frac{1}{K}\right) \bar{\mu}^\top \Sigma \bar{\mu} + \frac{\operatorname{tr}(\Sigma^2)}{MK}. \tag{24}$$

Cantelli's inequality applied to these exact moments yields the finite-sample alignment bound in Theorem 1. The two relative-variance terms motivate $d_\parallel$ and d_2, while $d_{\text{align}} = \min\{d_\parallel, d_2\}$ provides a conservative one-scalar summary: replacing both denominators by their minimum can only enlarge the upper bound. Combined with equation 21 and the local margin condition, this gives Theorem 2. Strict test-loss improvement additionally requires $\Delta_R^{\text{test}} + \Delta_{\text{ERM}} > 2(\epsilon_M + \epsilon_K)$. Positive alignment alone certifies the sign of the local margin, not a lower bound on its magnitude. The alignment bound itself needs only the second-moment assumptions of Theorem 1; the inherited uniform-generalization and residual-realizability conditions are needed only for the subsequent risk statement. For a different architecture or an unbounded loss, the radii ϵ_M, ϵ_K must be justified by a uniform-generalization result applicable to that setting; the alignment calculation does not itself provide them.

H.1 From Prior Width Bounds to an Instance-Specific Certificate

Cheng et al. (2026) established the function-preserving residual-expansion framework used in this paper. Their analysis connects an empirical jumpboard to the final expanded model through a train–test risk comparison, and bounds the probability that training- and test-averaged activation gradients are misaligned. The finite-sample alignment bound uses a covariance-spectral upper bound; its explicit width rate further requires control of total covariance energy and growth of the population gradient signal. It therefore characterizes a particular width-scaling regime, rather than the alignment reliability of a given model at an arbitrary insertion width.

We address the latter question by analyzing the alignment statistic directly. For a fixed reference model and insertion point, Theorem 1 gives

$$\operatorname{Var}(\mu_M^\top g_K) = \left(\frac{1}{M} + \frac{1}{K}\right) \bar{\mu}^\top \Sigma \bar{\mu} + \frac{\operatorname{tr}(\Sigma^2)}{MK}, \qquad \Pr(\mu_M^\top g_K \le 0) \le \frac{\Xi_{M,K}}{1 + \Xi_{M,K}}.$$

This fixed-width result requires independent sampling, $\mathbb{E}\|q\|^2 < \infty$, and $\bar{\mu} \neq 0$. It needs neither a prescribed width-uniform spectral bound nor an assumed growth law for $\bar{\mu}(N)$ or $\Sigma(N)$. Moreover, it retains the directional noise $\bar{\mu}^\top \Sigma \bar{\mu}$ and covariance energy $\operatorname{tr}(\Sigma^2)$ instead of replacing both by a worst-case spectral envelope. The resulting certificate can therefore describe anisotropic or irregular gradient geometry outside the prior width-growth regime; its numerical strength still depends on the actual signal–noise ratios.

The effective alignment dimension $d_{\text{align}} = \min\{d_\parallel, d_2\}$ compresses these two ratios into a conservative scalar certificate. Under the stronger growth conditions used in the prior width analysis,

$d_{\mathrm{align}}(N) = \Omega(N)$ and our bound recovers its width dependence. Without those conditions, Theorem 1 remains valid at each fixed width, while improvement as N increases is conditional on the behavior of $d_{\mathrm{align}}(N)$.

Our extension replaces the alignment-probability step of the earlier proof chain. The jumpboard construction, residual-realizability and model-selection conditions, and uniform-generalization argument remain inherited. In particular, finite second moments alone do not imply the generalization radii in Assumption 3: a conclusion about strict test-loss improvement of the final model still requires the margin condition stated in Theorem 2.

I Proof of Theorem 1

Proof. Define the centered empirical activation-gradient errors

$$\varepsilon_M := \mu_M - \bar{\mu}, \qquad \widetilde{\varepsilon}_K := g_K - \bar{\mu}.$$

Because the training and evaluation samples are drawn independently, ε_M and $\widetilde{\varepsilon}_K$ are independent. Moreover,

$$\mathbb{E}[\varepsilon_M] = \mathbb{E}[\widetilde{\varepsilon}_K] = 0,$$

and

$$\mathbb{E}\left[\varepsilon_M \varepsilon_M^\top\right] = \frac{\Sigma}{M}, \qquad \mathbb{E}\left[\widetilde{\varepsilon}_K \widetilde{\varepsilon}_K^\top\right] = \frac{\Sigma}{K}.$$

Let

$$X_{M,K} := \mu_M^\top g_K.$$

Using

$$\mu_M = \bar{\mu} + \varepsilon_M, \qquad g_K = \bar{\mu} + \widetilde{\varepsilon}_K,$$

we obtain

$$X_{M,K} = \|\bar{\mu}\|_2^2 + \bar{\mu}^\top \varepsilon_M + \bar{\mu}^\top \widetilde{\varepsilon}_K + \varepsilon_M^\top \widetilde{\varepsilon}_K.$$

The three random terms have zero expectation. In particular, independence and centering imply

$$\mathbb{E}\left[\varepsilon_M^\top \widetilde{\varepsilon}_K\right] = \mathbb{E}[\varepsilon_M]^\top \mathbb{E}[\widetilde{\varepsilon}_K] = 0.$$

Therefore,

$$\mathbb{E}[X_{M,K}] = \|\bar{\mu}\|_2^2.$$

We next compute the variance. For the first linear term,

$$\mathbb{E}\left[\left(\bar{\mu}^\top \varepsilon_M\right)^2\right] = \bar{\mu}^\top \mathbb{E}\left[\varepsilon_M \varepsilon_M^\top\right] \bar{\mu} = \frac{1}{M} \bar{\mu}^\top \Sigma \bar{\mu}.$$

Similarly,

$$\mathbb{E}\left[\left(\bar{\mu}^\top \widetilde{\varepsilon}_K\right)^2\right] = \frac{1}{K} \bar{\mu}^\top \Sigma \bar{\mu}.$$

For the bilinear term, independence gives

$$\begin{aligned} \mathbb{E}\left[\left(\varepsilon_M^\top \widetilde{\varepsilon}_K\right)^2\right] &= \mathbb{E}\left[\operatorname{tr}\left(\varepsilon_M \varepsilon_M^\top \widetilde{\varepsilon}_K \widetilde{\varepsilon}_K^\top\right)\right] \\ &= \operatorname{tr}\left(\mathbb{E}\left[\varepsilon_M \varepsilon_M^\top\right] \mathbb{E}\left[\widetilde{\varepsilon}_K \widetilde{\varepsilon}_K^\top\right]\right) \\ &= \frac{1}{MK} \operatorname{tr}(\Sigma^2). \end{aligned} \tag{25}$$

All cross terms vanish. First,

$$\mathbb{E}\left[\left(\bar{\mu}^\top \varepsilon_M\right)\left(\bar{\mu}^\top \widetilde{\varepsilon}_K\right)\right] = 0$$

by independence and centering. Furthermore,

$$\begin{aligned} \mathbb{E}\left[\left(\bar{\mu}^\top \varepsilon_M\right)\left(\varepsilon_M^\top \widetilde{\varepsilon}_K\right)\right] &= \mathbb{E}_{\varepsilon_M}\left[\left(\bar{\mu}^\top \varepsilon_M\right) \varepsilon_M^\top \mathbb{E}_{\widetilde{\varepsilon}_K}\left[\widetilde{\varepsilon}_K\right]\right] \\ &= 0. \end{aligned} \tag{26}$$

The analogous cross term involving $\bar{\mu}^\top \widetilde{\varepsilon}_K$ also vanishes. Hence

$$\mathrm{Var}(X_{M,K}) = \left(\frac{1}{M} + \frac{1}{K}\right) \bar{\mu}^\top \Sigma \bar{\mu} + \frac{1}{MK} \operatorname{tr}(\Sigma^2).$$

Set

$$m := \mathbb{E}[X_{M,K}] = \|\bar{\mu}\|_2^2 > 0, \qquad v := \mathrm{Var}(X_{M,K}).$$

The activation-gradient misalignment event satisfies

$$\{X_{M,K} \le 0\} = \{X_{M,K} - m \le -m\}.$$

By Cantelli's one-sided inequality,

$$\Pr(X_{M,K} \le 0) \le \frac{v}{v + m^2}.$$

Since $m^2 = \|\bar{\mu}\|_2^4$, we have

$$\begin{aligned} \frac{v}{m^2} &= \left(\frac{1}{M} + \frac{1}{K}\right) \frac{\bar{\mu}^\top \Sigma \bar{\mu}}{\|\bar{\mu}\|_2^4} + \frac{1}{MK} \frac{\operatorname{tr}(\Sigma^2)}{\|\bar{\mu}\|_2^4} \\ &= \frac{1}{M d_\parallel} + \frac{1}{K d_\parallel} + \frac{1}{MK d_2} \\ &= \Xi_{M,K}. \end{aligned} \tag{27}$$

Therefore,

$$\Pr\big(\mu_M^\top g_K \le 0\big) \le \frac{\Xi_{M,K}}{1 + \Xi_{M,K}}.$$

Finally, because

$$d_{\mathrm{align}} = \min\{d_\parallel, d_2\},$$

we have

$$\frac{1}{d_\parallel} \le \frac{1}{d_{\mathrm{align}}}, \qquad \frac{1}{d_2} \le \frac{1}{d_{\mathrm{align}}}.$$

It follows that

$$\begin{aligned} \Xi_{M,K} &\le \left(\frac{1}{M} + \frac{1}{K} + \frac{1}{MK}\right) \frac{1}{d_{\mathrm{align}}} \\ &= \frac{M + K + 1}{MK\, d_{\mathrm{align}}} \\ &= \frac{1}{n_{\mathrm{eff}} d_{\mathrm{align}}}, \end{aligned} \tag{28}$$

where

$$n_{\mathrm{eff}} = \frac{MK}{M + K + 1}.$$

Since $x \mapsto x/(1+x)$ is nondecreasing on $[0, \infty)$,

$$\frac{\Xi_{M,K}}{1 + \Xi_{M,K}} \le \frac{1}{1 + n_{\mathrm{eff}} d_{\mathrm{align}}}.$$

This completes the proof. □

Corollary 3 (Balanced sample allocation)**.** *Fix an integer sampling budget $B \ge 2$ and require*

$$M + K = B, \qquad M, K \in \mathbb{N}_+.$$

Then the effective sample size

$$n_{\mathrm{eff}}(M, K) = \frac{MK}{M + K + 1} = \frac{MK}{B + 1}$$

is maximized when the training and test sample sizes are as balanced as possible, namely,

$$|M - K| \le 1.$$

More precisely, an optimal allocation is

$$M^\star = \left\lfloor \frac{B}{2} \right\rfloor, \qquad K^\star = \left\lceil \frac{B}{2} \right\rceil,$$

or the reversed allocation. The corresponding maximum effective sample size is

$$n_{\text{eff}}^\star = \frac{\lfloor B^2/4 \rfloor}{B+1}.$$

Consequently, for fixed B and d_{align}, the simplified activation-gradient misalignment bound

$$\Pr\big(\mu_M^\top g_K \le 0\big) \le \frac{1}{1 + n_{\text{eff}}(M,K) d_{\text{align}}}$$

is minimized by a balanced allocation of the sampling budget.

Proof. Because $M + K = B$ is fixed, maximizing $n_{\text{eff}}(M, K)$ is equivalent to maximizing the product MK. Using

$$(M-K)^2 \ge 0,$$

we obtain

$$(M+K)^2 \ge 4MK,$$

and hence

$$MK \le \frac{B^2}{4}.$$

For positive integer M and K, the largest attainable value is

$$MK = \left\lfloor \frac{B^2}{4} \right\rfloor,$$

which is achieved exactly when M and K differ by at most one. Therefore,

$$n_{\text{eff}}(M,K) \le \frac{\lfloor B^2/4 \rfloor}{B+1}.$$

Finally, the function

$$x \longmapsto \frac{1}{1 + x\, d_{\text{align}}}$$

is nonincreasing for $x \ge 0$. Thus, maximizing $n_{\text{eff}}(M, K)$ minimizes the corresponding simplified misalignment-probability bound. □

J Proof of Theorem 2

Proof. **Fixed-jumpboard comparison.** Let $\widetilde{f}_S$ denote the empirical jumpboard model in Theorem 2, and let $\Delta_R^{\text{test}} := L_{\text{test}}(f^*_{\text{old}}) - L_{\text{test}}(\widetilde{f}_S)$ denote its associated finite-test margin. Define the two uniform-generalization events

$$\mathcal{E}_M := \left\{ \sup_{f \in \mathcal{H}_{\text{new}}} |R(f) - L_{\text{train}}(f)| \le \epsilon_M \right\},$$

and

$$\mathcal{E}_K := \left\{ \sup_{f \in \mathcal{H}_{\text{new}}} |R(f) - L_{\text{test}}(f)| \le \epsilon_K \right\}.$$

By Assumption 3,

$$\Pr(\mathcal{E}_M) \ge 1 - \delta, \qquad \Pr(\mathcal{E}_K) \ge 1 - \delta.$$

Therefore, by the union bound,

$$\Pr(\mathcal{E}_M \cap \mathcal{E}_K) \ge 1 - 2\delta.$$

No independence between $\mathcal{E}_M$ and $\mathcal{E}_K$ is required.

We now work on the event $\mathcal{E}_M \cap \mathcal{E}_K$. Because $f_{\mathrm{new}} \in \mathcal{H}_{\mathrm{new}}$, the test-side and training-side uniform bounds imply

$$\begin{aligned} L_{\mathrm{test}}(f_{\mathrm{new}}) &\le R(f_{\mathrm{new}}) + \epsilon_K \\ &\le L_{\mathrm{train}}(f_{\mathrm{new}}) + \epsilon_M + \epsilon_K. \end{aligned} \tag{29}$$

By the definition of the optimization gain,

$$\Delta_{\mathrm{ERM}} = L_{\mathrm{train}}(\widetilde{f}_S) - L_{\mathrm{train}}(f_{\mathrm{new}}),$$

and hence

$$L_{\mathrm{train}}(f_{\mathrm{new}}) = L_{\mathrm{train}}(\widetilde{f}_S) - \Delta_{\mathrm{ERM}}.$$

Substituting this identity into equation 29 gives

$$L_{\mathrm{test}}(f_{\mathrm{new}}) \le L_{\mathrm{train}}(\widetilde{f}_S) - \Delta_{\mathrm{ERM}} + \epsilon_M + \epsilon_K. \tag{30}$$

Since $\widetilde{f}_S \in \mathcal{H}_{\mathrm{new}}$, the same uniform events yield

$$L_{\mathrm{train}}(\widetilde{f}_S) \le R(\widetilde{f}_S) + \epsilon_M \le L_{\mathrm{test}}(\widetilde{f}_S) + \epsilon_M + \epsilon_K.$$

Combining this inequality with equation 30, we obtain

$$L_{\mathrm{test}}(f_{\mathrm{new}}) \le L_{\mathrm{test}}(\widetilde{f}_S) - \Delta_{\mathrm{ERM}} + 2(\epsilon_M + \epsilon_K). \tag{31}$$

By the definition of the finite-test jumpboard margin,

$$\Delta_R^{\mathrm{test}} = L_{\mathrm{test}}(f^*_{\mathrm{old}}) - L_{\mathrm{test}}(\widetilde{f}_S),$$

so that

$$L_{\mathrm{test}}(\widetilde{f}_S) = L_{\mathrm{test}}(f^*_{\mathrm{old}}) - \Delta_R^{\mathrm{test}}.$$

Substitution into equation 31 therefore gives

$$L_{\mathrm{test}}(f_{\mathrm{new}}) \le L_{\mathrm{test}}(f^*_{\mathrm{old}}) - \Delta_R^{\mathrm{test}} - \Delta_{\mathrm{ERM}} + 2(\epsilon_M + \epsilon_K). \tag{32}$$

This inequality holds with probability at least $1 - 2\delta$.

Consequently, on the same event, whenever

$$\Delta_R^{\mathrm{test}} + \Delta_{\mathrm{ERM}} > 2(\epsilon_M + \epsilon_K),$$

we have

$$L_{\mathrm{test}}(f_{\mathrm{new}}) < L_{\mathrm{test}}(f^*_{\mathrm{old}}),$$

which proves the strict finite-test risk improvement statement and the first part of Theorem 2.

Alignment implies a positive margin for the selected jumpboard. We now prove the additional alignment-supported statement. Under its additional premise, the empirical jumpboard $\widetilde{f}_S$ is selected along the differentiable function-preserving residual-expansion path from Appendix G. We therefore write

$$\widetilde{f}_S = \widetilde{f}_{S_{\mathrm{train}},\eta_S}$$

for some selected step size $\eta_S > 0$, with $\widetilde{f}_{S_{\mathrm{train}},0} = f^*_{\mathrm{old}}$, and define the path-indexed finite-test margin by

$$\Delta_R^{\mathrm{test}}(\eta) := L_{\mathrm{test}}(f^*_{\mathrm{old}}) - L_{\mathrm{test}}\left(\widetilde{f}_{S_{\mathrm{train}},\eta}\right). \tag{33}$$

For the selected jumpboard,

$$\Delta_R^{\mathrm{test}} = \Delta_R^{\mathrm{test}}(\eta_S). \tag{34}$$

Under the residual-realizability condition stated in Appendix G, the first-order residual perturbation induced by the training sample satisfies

$$\frac{1}{K}\sum_{j=1}^{K} \widetilde{q}_j^{\top} u_{S_{\mathrm{train}}}(\widetilde{z}_j) = -\mu_M^{\top} g_K.$$

Applying the chain rule at the function-preserving initialization therefore gives

$$\frac{d}{d\eta} L_{\text{test}}\left(\widetilde{f}_{S_{\text{train}},\eta}\right)\Big|_{\eta=0} = -\mu_M^\top g_K. \tag{35}$$

Equivalently, differentiability at $\eta = 0$ yields the local expansion

$$L_{\text{test}}\left(\widetilde{f}_{S_{\text{train}},\eta}\right) = L_{\text{test}}(f^*_{\text{old}}) - \eta\mu_M^\top g_K + o(\eta). \tag{36}$$

Using equation 33, this is equivalently

$$\Delta_R^{\text{test}}(\eta) = \eta\mu_M^\top g_K + o(\eta). \tag{37}$$

Consider any realization for which

$$\mu_M^\top g_K > 0. \tag{38}$$

Since the remainder in equation 37 satisfies $o(\eta)/\eta \to 0$ as $\eta \downarrow 0$, there exists $\eta_0 > 0$ such that, for every $0 < \eta \le \eta_0$,

$$|o(\eta)| < \frac{1}{2}\eta\mu_M^\top g_K. \tag{39}$$

Combining equation 37 and equation 39 gives

$$\Delta_R^{\text{test}}(\eta) > \frac{1}{2}\eta\mu_M^\top g_K > 0, \qquad 0 < \eta \le \eta_0. \tag{40}$$

The "small-step regime" premise in Theorem 2 means that, whenever the first-order coefficient is positive, the selected empirical jumpboard step lies inside this local regime, namely $0 < \eta_S \le \eta_0$. Therefore, equation 34 and equation 40 imply

$$\mu_M^\top g_K > 0 \quad \Longrightarrow \quad \Delta_R^{\text{test}} > 0. \tag{41}$$

Probability of a positive jumpboard margin. Define the positive-alignment event and the positive-margin event by

$$A_{M,K} := \left\{\mu_M^\top g_K > 0\right\}, \qquad F_{M,K} := \left\{\Delta_R^{\text{test}} > 0\right\}.$$

The implication in equation 41 gives

$$A_{M,K} \subseteq F_{M,K}.$$

Hence

$$\Pr(F_{M,K}) \ge \Pr(A_{M,K}) = 1 - \Pr\left(\mu_M^\top g_K \le 0\right). \tag{42}$$

Applying Theorem 1 to equation 42 yields

$$\Pr\left(\Delta_R^{\text{test}} > 0\right) \ge 1 - B_{\text{exact}}(M,K) \ge 1 - B_{\text{align}}(M,K). \tag{43}$$

This proves the positive-margin statement in Theorem 2.

Joint high-probability statement. Finally, consider the three events $\mathcal{E}_M$, $\mathcal{E}_K$, and $F_{M,K}$. No independence among these events is required. By Assumption 3 and equation 43,

$$\Pr(\mathcal{E}_M^c) \le \delta, \qquad \Pr(\mathcal{E}_K^c) \le \delta, \qquad \Pr(F_{M,K}^c) \le B_{\text{exact}}(M,K).$$

The union bound therefore gives

$$\Pr(\mathcal{E}_M \cap \mathcal{E}_K \cap F_{M,K}) \ge 1 - 2\delta - B_{\text{exact}}(M,K). \tag{44}$$

Since $B_{\text{exact}}(M,K) \le B_{\text{align}}(M,K)$, The bound in equation 44 further implies

$$\Pr(\mathcal{E}_M \cap \mathcal{E}_K \cap F_{M,K}) \ge 1 - 2\delta - B_{\text{align}}(M,K). \tag{45}$$

The event $F_{M,K}$ guarantees only that the selected jumpboard has a positive finite-test margin; Theorem 1 does not provide a lower bound on the magnitude of that margin. Therefore, even on the joint event characterized by equation 44 and equation 45, strict improvement of the final expanded model continues to require the realized condition

$$\Delta_R^{\text{test}} + \Delta_{\text{ERM}} > 2(\epsilon_M + \epsilon_K).$$

Together with equation 32, this completes the proof. □

K Recovery of the Previous Width-Dependent Rate

The finite-sample certificate in Theorem 1 does not require any prescribed relation between the activation-gradient distribution and the model width. We now show that the stronger covariance-spectrum, total-variance, and signal-growth conditions used in the previous width-dependent analysis imply linear growth of the effective alignment dimension and therefore recover the same-order width-dependent misalignment rate.

Proposition 4 (Recovery of the width-dependent rate)**.** *For each insertion width N, let $\bar{\mu}^{(N)}$ and $\Sigma^{(N)}$ denote the population mean and covariance of the corresponding activation-gradient signal. Suppose that there exist constants $c_\mu > 0$, $C_\lambda > 0$, $C_{\mathrm{tr}} > 0$, and $N_0 \in \mathbb{N}$, all independent of N, together with a positive variance scale τ_N^2, such that for every $N \geq N_0$,*

$$\left\|\bar{\mu}^{(N)}\right\|_2^2 \geq c_\mu N \tau_N^2, \tag{46}$$

$$\lambda_{\max}\left(\Sigma^{(N)}\right) \leq C_\lambda \tau_N^2, \tag{47}$$

$$\mathrm{tr}\left(\Sigma^{(N)}\right) \leq C_{\mathrm{tr}} N \tau_N^2. \tag{48}$$

Then, for every $N \geq N_0$,

$$d_{\|}(N) \geq \frac{c_\mu}{C_\lambda} N, \tag{49}$$

$$d_2(N) \geq \frac{c_\mu^2}{C_\lambda C_{\mathrm{tr}}} N. \tag{50}$$

Consequently,

$$d_{\mathrm{align}}(N) \geq c_0 N,$$

where

$$c_0 := \min\left\{\frac{c_\mu}{C_\lambda}, \frac{c_\mu^2}{C_\lambda C_{\mathrm{tr}}}\right\} > 0.$$

Therefore, for every $M, K \in \mathbb{N}_+$,

$$\begin{aligned}\Pr\left(\mu_M^\top g_K \leq 0\right) &\leq B_{\mathrm{align}}(M, K; N) \\ &\leq \frac{1}{1 + c_0 n_{\mathrm{eff}}(M, K) N}.\end{aligned} \tag{51}$$

In particular,

$$\Pr\left(\mu_M^\top g_K \leq 0\right) = O\left(\frac{1}{MN} + \frac{1}{KN} + \frac{1}{MKN}\right).$$

Proof. We first lower-bound the directional effective dimension. Since $\Sigma^{(N)}$ is positive semidefinite,

$$\bar{\mu}^{(N)\top} \Sigma^{(N)} \bar{\mu}^{(N)} \leq \lambda_{\max}\left(\Sigma^{(N)}\right) \left\|\bar{\mu}^{(N)}\right\|_2^2.$$

Therefore,

$$\begin{aligned}d_{\|}(N) &= \frac{\left\|\bar{\mu}^{(N)}\right\|_2^4}{\bar{\mu}^{(N)\top} \Sigma^{(N)} \bar{\mu}^{(N)}} \\ &\geq \frac{\left\|\bar{\mu}^{(N)}\right\|_2^2}{\lambda_{\max}\left(\Sigma^{(N)}\right)}.\end{aligned} \tag{52}$$

Applying equation 46 and equation 47 gives

$$d_{\|}(N) \geq \frac{c_\mu}{C_\lambda} N.$$

We next lower-bound the covariance-energy effective dimension. Let $\lambda_1^{(N)}, \dots, \lambda_N^{(N)}$ denote the eigenvalues of $\Sigma^{(N)}$. Because these eigenvalues are nonnegative,

$$\begin{aligned}\operatorname{tr}\left(\Sigma^{(N)2}\right) &= \sum_{i=1}^{N} \left(\lambda_i^{(N)}\right)^2 \\ &\le \lambda_{\max}\left(\Sigma^{(N)}\right) \sum_{i=1}^{N} \lambda_i^{(N)} \\ &= \lambda_{\max}\left(\Sigma^{(N)}\right) \operatorname{tr}\left(\Sigma^{(N)}\right). \end{aligned} \tag{53}$$

Using equation 47 and equation 48, we obtain

$$\operatorname{tr}\left(\Sigma^{(N)2}\right) \le C_\lambda C_{\mathrm{tr}} N \tau_N^4.$$

Hence,

$$\begin{aligned} d_2(N) &= \frac{\left\|\bar{\mu}^{(N)}\right\|_2^4}{\operatorname{tr}\left(\Sigma^{(N)2}\right)} \\ &\ge \frac{c_\mu^2 N^2 \tau_N^4}{C_\lambda C_{\mathrm{tr}} N \tau_N^4} \\ &= \frac{c_\mu^2}{C_\lambda C_{\mathrm{tr}}} N. \end{aligned} \tag{54}$$

Because

$$d_{\mathrm{align}}(N) = \min\left\{d_{\|}(N), d_2(N)\right\},$$

the two preceding bounds imply

$$d_{\mathrm{align}}(N) \ge c_0 N,$$

with

$$c_0 = \min\left\{\frac{c_\mu}{C_\lambda}, \frac{c_\mu^2}{C_\lambda C_{\mathrm{tr}}}\right\}.$$

Applying the simplified certificate from Theorem 1 yields

$$\begin{aligned} \Pr\left(\mu_M^\top g_K \le 0\right) &\le \frac{1}{1 + n_{\mathrm{eff}}(M,K) d_{\mathrm{align}}(N)} \\ &\le \frac{1}{1 + c_0 n_{\mathrm{eff}}(M,K) N}. \end{aligned} \tag{55}$$

Since

$$n_{\mathrm{eff}}(M,K) = \frac{MK}{M+K+1},$$

we have

$$\begin{aligned} \frac{1}{c_0 n_{\mathrm{eff}}(M,K) N} &= \frac{M+K+1}{c_0 MKN} \\ &= \frac{1}{c_0 KN} + \frac{1}{c_0 MN} + \frac{1}{c_0 MKN}. \end{aligned} \tag{56}$$

Therefore,

$$\Pr\left(\mu_M^\top g_K \le 0\right) = O\left(\frac{1}{MN} + \frac{1}{KN} + \frac{1}{MKN}\right).$$

This completes the proof. □

The same rate can also be recovered directly from the exact certificate. Under the conditions of Proposition 4,

$$\begin{aligned}\Xi_{M,K}(N) &= \frac{1}{Md_{\|}(N)} + \frac{1}{Kd_{\|}(N)} + \frac{1}{MKd_2(N)} \\ &\le \frac{C_\lambda}{c_\mu N}\left(\frac{1}{M}+\frac{1}{K}\right) + \frac{C_\lambda C_{\mathrm{tr}}}{c_\mu^2 MKN}.\end{aligned} \tag{57}$$

Since

$$B_{\mathrm{exact}}(M,K;N) = \frac{\Xi_{M,K}(N)}{1+\Xi_{M,K}(N)} \le \Xi_{M,K}(N),$$

the exact certificate satisfies the same asymptotic rate.

These conditions are sufficient rather than necessary. Theorem 1 remains valid even when the signal norm, covariance spectrum, or total variance does not obey the prescribed width-growth relations. Their role is only to provide one particular route to $d_{\mathrm{align}}(N) = \Omega(N)$ and thereby recover the previous width-dependent guarantee.

L Exact Parameter-Space Realization of the Activation-Gradient Direction

We now show that the activation-gradient direction used in the finite-sample analysis can be realized exactly by a concrete function-preserving residual module. All quantities z, q, μ_M, g_K, L_{train}, and L_{test} follow the notation introduced above.

Consider an inserted residual module of the form

$$h_{U,V,b}(z) = V\psi_U(z) + b,$$

where $V \in \mathbb{R}^{N\times m}$ is the output projection and $b \in \mathbb{R}^N$ is an additive output parameter. Initialize the new module at

$$U = U_0, \qquad V = 0, \qquad b = 0.$$

Because

$$h_{U_0,0,0}(z) = 0 \qquad \text{for every } z,$$

the insertion preserves the original network function exactly.

At this function-preserving initialization, the sample-level gradients with respect to the new output parameters satisfy

$$\begin{aligned}\nabla_V \ell\big(f_{\mathrm{top}}\big(z + V\psi_{U_0}(z) + b\big), y\big)\big|_{V=0,b=0} &= q(x,y)\psi_{U_0}(z)^\top, \\ \nabla_b \ell\big(f_{\mathrm{top}}\big(z + V\psi_{U_0}(z) + b\big), y\big)\big|_{V=0,b=0} &= q(x,y).\end{aligned}$$

Thus, $q(x,y)$ is the prediction-side factor of the output-projection weight gradient and, more importantly, is exactly the sample-level parameter gradient with respect to the additive output parameter b.

Proposition 5 (Exact realization of the activation-gradient direction)**.** *Suppose that the inserted residual module contains the zero-initialized additive output parameter b. Then the training-selected direction $-\mu_M$ is realized by the parameter path*

$$U(\eta) = U_0, \qquad V(\eta) = 0, \qquad b(\eta) = -\eta\mu_M.$$

The induced residual perturbation satisfies

$$u_{S_{\mathrm{train}}}(z) = -\mu_M \qquad \textit{for every } z,$$

and the directional derivative of the independent finite-test risk is

$$\frac{d}{d\eta} L_{\mathrm{test}}\big(\widetilde{f}_{S_{\mathrm{train}},\eta}\big)\Big|_{\eta=0} = -\mu_M^\top g_K. \tag{58}$$

Consequently, positive train–test activation-gradient alignment implies that a sufficiently small realizable update of the inserted residual module strictly decreases the independent finite-test risk.

Proof. Along the stated parameter path,

$$h_\eta(z) = -\eta\mu_M,$$

and hence

$$u_{S_{\text{train}}}(z) = \frac{\partial}{\partial\eta} h_\eta(z)\Big|_{\eta=0} = -\mu_M.$$

Therefore, the residual-realizability condition equation 15 holds exactly.

Applying the chain rule to the empirical training risk gives

$$\frac{d}{d\eta} L_{\text{train}}\big(\widetilde{f}_{S_{\text{train}},\eta}\big)\Big|_{\eta=0} = -\|\mu_M\|_2^2.$$

Thus, whenever $\mu_M \neq 0$, the path is a strict first-order descent direction for the empirical training risk.

For the independent test sample, the same calculation gives

$$\begin{aligned}\frac{d}{d\eta} L_{\text{test}}\big(\widetilde{f}_{S_{\text{train}},\eta}\big)\Big|_{\eta=0} &= \frac{1}{K}\sum_{j=1}^{K} \widetilde{q}_j^{\top}(-\mu_M) \\ &= -\mu_M^{\top} g_K,\end{aligned}$$

which proves equation 58.

If $\mu_M^{\top} g_K > 0$, the derivative is strictly negative. By differentiability at $\eta = 0$,

$$L_{\text{test}}\big(\widetilde{f}_{S_{\text{train}},\eta}\big) = L_{\text{test}}(f) - \eta\mu_M^{\top} g_K + o(\eta).$$

Hence, there exists $\eta_0 > 0$ such that

$$L_{\text{test}}\big(\widetilde{f}_{S_{\text{train}},\eta}\big) < L_{\text{test}}(f)$$

for every $0 < \eta \leq \eta_0$. □

The additive output parameter has an equivalent constant-channel representation. Specifically, define

$$\widetilde{\psi}_U(z) = \begin{bmatrix}\psi_U(z) \\ 1\end{bmatrix}, \qquad \widetilde{V} = [V \quad b].$$

Then

$$V\psi_U(z) + b = \widetilde{V}\widetilde{\psi}_U(z).$$

Thus, an affine output projection, a zero-initialized output bias, and a constant residual channel provide equivalent realizations of the same activation-space tangent direction. This requirement applies only to the newly inserted module and does not require the pretrained base architecture to use biased linear layers.

If the inserted module contains neither an additive output parameter nor a constant feature, $q(x, y)$ alone does not generally determine a realizable parameter-space update. In that case, the appropriate sample-level signal is the full output-projection gradient

$$\zeta(x, y) = \text{vec}\left(q(x, y)\psi_{U_0}(z)^{\top}\right),$$

and the finite-sample alignment analysis must instead be formulated in terms of ζ.

M Architecture-Specific Selection of Activation-Gradient Measurement Locations

The normalized residual-expansion framework is formulated at a designated intermediate representation and therefore permits architecture-specific measurement locations. The selected location must provide a valid residual insertion point, preserve the function-preserving expansion interpretation, and allow the activation gradient to represent the relevant first-order improvement signal. Based on these principles, we use the final residual representation for Transformer language models and an early residual block for ResNet-20.

M.1 EXISTENTIAL NATURE OF THE LAYER-WISE THEORY.

Two distinct statements are existential over admissible insertion locations: the population residual-expansion condition and the finite-test alignment certificate. They must not be identified with one another. Let $\mathcal{L}_{\mathrm{adm}}$ denote the set of admissible insertion layers. For each $\ell \in \mathcal{L}_{\mathrm{adm}}$, the corresponding gradient signal, mean, covariance, effective alignment dimension, and misalignment probability are defined locally at that layer.

For the population statement, it is sufficient that

$$\exists\, \ell_{\mathrm{pop}} \in \mathcal{L}_{\mathrm{adm}} \quad \text{such that} \quad \mathcal{C}_{\ell_{\mathrm{pop}}} \text{ holds}, \tag{59}$$

where $\mathcal{C}_\ell$ is the population expansion condition. The population-level residual-expansion argument applied at ℓ_{pop} then supplies an improving population-risk perturbation. This conclusion is expressed in terms of R and does not follow from, or make a claim about, a finite held-out test sample.

Separately, for the finite-test statement, fix a layer ℓ_{align} independently of the realized training and test samples and suppose that

$$B_{M,K}^{(\ell_{\mathrm{align}})} \le \delta, \tag{60}$$

where $B_{M,K}^{(\ell)}$ is the finite-sample upper bound on gradient misalignment at layer ℓ. Theorem 1 then gives

$$\Pr\left[\left(\mu_M^{(\ell_{\mathrm{align}})}\right)^\top g_K^{(\ell_{\mathrm{align}})} > 0\right] \ge 1 - \delta. \tag{61}$$

The witness layer ℓ_{align} may be specified from population moments $\bar{\mu}_\ell, \Sigma_\ell$ or by architectural convention, but it is *not* selected from the realized finite sample. Thus, Equation 61 is the single-layer bound of Theorem 1 at a pre-fixed insertion point, with no multiple-testing correction. Were ℓ_{align} instead chosen in a data-dependent manner (e.g. the layer with the largest empirical $\hat{d}_{\mathrm{align}}$), a union bound over $\mathcal{L}_{\mathrm{adm}}$ would be required.

Equation 61 is only a finite-sample statement: it says that a training-estimated direction has positive inner product with the independently estimated test direction. It neither estimates the population directional derivative nor implies lower population risk. Likewise, the population conclusion from Equation 59 does not imply finite-test alignment. The two witness layers may coincide, but this is not required; even when they do coincide, the conclusions remain logically separate and need not concern the same residual perturbation. Each statement requires a valid layer, rather than validity at every admissible layer, and failure of one statement does not establish failure of the other.

M.2 WHY THE FINAL RESIDUAL REPRESENTATION IS MEASURED IN LLMS

The original normalized residual-expansion framework is defined at an arbitrary designated intermediate representation. Specifically, a trained network can be decomposed as $f = f_{\mathrm{top}} \circ f_{\mathrm{bot}}$, where $z = f_{\mathrm{bot}}(x) \in \mathbb{R}^N$ denotes the residual-stream representation at the selected insertion point and $q(x, y) = \nabla_z \ell(f_{\mathrm{top}}(z), y)$ denotes the corresponding activation-gradient signal. The theory therefore does not prescribe a unique layer shared by all architectures. Instead, the selected location must constitute a valid residual insertion point and must preserve the interpretation of q as the first-order signal governing a function-preserving residual expansion.

Consider an L-block Transformer with residual-stream representations $z_0, z_1, \ldots, z_L$, where

$$z_{l+1} = T_l(z_l), \qquad l = 0, \ldots, L-1,$$

and let H denote the final normalization and language-modeling head. For each residual-stream location l, define

$$q_l(x, y) = \nabla_{z_l} \ell(H(z_L), y)\,.$$

Let

$$\Phi_{l \to L} = T_{L-1} \circ \cdots \circ T_l$$

denote the downstream Transformer mapping from z_l to z_L, and define its Jacobian by

$$A_l(x) = J_{\Phi_{l \to L}}(z_l).$$

The chain rule then gives

$$q_l(x,y) = A_l(x)^\top q_L(x,y),$$

where

$$q_L(x,y) = \nabla_{z_L} \ell(H(z_L), y)\,.$$

Thus, an activation gradient measured at an earlier layer is obtained by transporting the terminal activation gradient through every remaining Transformer block.

This transport directly affects the population quantities entering the effective alignment dimension. At layer l, the population mean and covariance are

$$\bar{\mu}_l = \mathbb{E}\left[A_l(x)^\top q_L(x,y)\right]$$

and

$$\Sigma_l = \mathrm{Cov}\left(A_l(x)^\top q_L(x,y)\right).$$

Consequently, $\bar{\mu}_l$ and Σ_l depend not only on the terminal loss-gradient geometry, but also on the anisotropy, conditioning, sample dependence, and gradient correlation induced by the downstream Jacobian $A_l(x)$. The corresponding effective dimensions

$$d_{\|}^{(l)} = \frac{\|\bar{\mu}_l\|_2^4}{\bar{\mu}_l^\top \Sigma_l \bar{\mu}_l}, \qquad d_2^{(l)} = \frac{\|\bar{\mu}_l\|_2^4}{\mathrm{tr}(\Sigma_l^2)}$$

therefore combine the intrinsic activation-gradient geometry at the prediction interface with an additional layer-dependent transport effect.

The distinction can be made explicit by first considering an idealized deterministic downstream Jacobian of the form

$$A_l = cR,$$

where $c > 0$ and $R \in \mathbb{R}^{N\times N}$ is orthogonal:

$$R^\top R = RR^\top = I.$$

In this case,

$$\bar{\mu}_l = cR^\top \bar{\mu}_L, \qquad \Sigma_l = c^2 R^\top \Sigma_L R.$$

Orthogonal invariance then gives

$$\|\bar{\mu}_l\|_2^4 = c^4 \|\bar{\mu}_L\|_2^4,$$
$$\bar{\mu}_l^\top \Sigma_l \bar{\mu}_l = c^4 \bar{\mu}_L^\top \Sigma_L \bar{\mu}_L,$$

and

$$\mathrm{tr}(\Sigma_l^2) = c^4\,\mathrm{tr}(\Sigma_L^2).$$

It follows that

$$d_{\|}^{(l)} = d_{\|}^{(L)}, \qquad d_2^{(l)} = d_2^{(L)}, \qquad d_{\mathrm{align}}^{(l)} = d_{\mathrm{align}}^{(L)}.$$

Hence, a deterministic scaled isometry preserves the effective alignment geometry exactly. Differences between early- and late-layer measurements arise from deviations from this ideal case, including unequal singular values, sample-dependent Jacobians, and correlations between the downstream transformation and the terminal gradient.

To isolate the sample-dependent contribution, write

$$A_l(x) = \overline{A}_l + \Delta A_l(x), \qquad \overline{A}_l = \mathbb{E}[A_l(x)].$$

The earlier-layer gradient can then be decomposed as

$$q_l = \overline{A}_l^\top q_L + r_l, \qquad r_l = \Delta A_l(x)^\top q_L.$$

The additional transport term satisfies

$$\mathbb{E}\|r_l\|_2^2 \le \mathbb{E}\left[\|\Delta A_l(x)\|_{\mathrm{op}}^2 \|q_L(x,y)\|_2^2\right].$$

Accordingly, input-dependent variation in the downstream Transformer Jacobian introduces an additional source of second-moment variation that can modify both the estimated signal direction and its covariance. Moreover, because

$$A_l(x) = J_{T_{L-1}}(z_{L-1}) \cdots J_{T_l}(z_l),$$

earlier residual representations involve a longer product of attention- and feed-forward-block Jacobians, providing more opportunities for anisotropic scaling, rotation, and sample-dependent distortion.

At the output of the final Transformer block, the downstream Transformer mapping is the identity:

$$A_L = I.$$

Therefore,

$$q_L = \nabla_{z_L} \ell(H(z_L), y)\,,$$

and the transport perturbation satisfies

$$\Delta A_L = 0, \qquad r_L = 0.$$

The final residual representation consequently removes all cumulative Jacobian transport through additional Transformer blocks. Its mean $\bar{\mu}_L$, covariance Σ_L, and effective alignment dimension $d_{\text{align}}^{(L)}$ provide the most direct characterization of the activation-gradient geometry connected to the language-modeling loss.

This location remains fully compatible with the residual-expansion mechanism of the original theory. A Transformer MLP branch can be written as

$$h_{U,V}(z) = W_{\text{out}} \psi_U(z),$$

while an attention branch can be written as

$$h_{U,V}(z) = W_O \,\text{Attn}_U(z).$$

In both cases, the output projection plays the role of V in the abstract residual form $h_{U,V}(z) = V\psi_U(z)$. Initializing this output projection at zero gives $h_{U,0}(z) = 0$, preserving the original model function while retaining a realizable first-order residual direction. The final residual stream is therefore an admissible insertion point under the same function-preserving residual-expansion framework used by the theory.

For the LLaMA-style and Pythia models, we consequently measure the activation gradient at the output of the final Transformer block and before the final normalization and language-modeling head. This choice follows a minimal-transport principle: among residual-stream locations, it eliminates the largest amount of downstream Jacobian-induced variation and provides the least confounded measurement of width-dependent activation-gradient geometry. We do not claim that the final block universally maximizes d_{align} over all possible layers; rather, it is the canonical location at which the measured statistics are not additionally transformed by any subsequent Transformer block.

M.3 Why an Early Residual Block Is Measured in ResNet

The original normalized residual-expansion framework permits the insertion point to be placed at an arbitrary designated intermediate representation. For a ResNet with intermediate representations $z_0, z_1, \ldots, z_L$, write

$$z_{j+1} = T_j(z_j), \qquad j = 0, \ldots, L-1,$$

where T_j denotes the j-th residual block, including the skip connection and the residual branch. At a candidate insertion point l, define the activation gradient

$$q_l(x, y) = \nabla_{z_l} \ell(f(x), y)\,.$$

A function-preserving residual expansion introduces a new branch $h_{l,\eta}$ satisfying $h_{l,0}(z) = 0$. Its first-order perturbation is

$$u_l(z_l) = \left.\frac{\partial}{\partial \eta} h_{l,\eta}(z_l)\right|_{\eta=0}.$$

The corresponding population-risk derivative is

$$\left.\frac{\mathrm{d}}{\mathrm{d}\eta} R(f_{l,\eta})\right|_{\eta=0} = \mathbb{E}\left[q_l(x,y)^\top u_l(z_l)\right].$$

Therefore, the suitability of an insertion point depends on whether the realizable residual tangent family contains a direction that is non-orthogonal to the activation-gradient signal.

Let $\mathcal{U}_l$ denote the set of first-order perturbations realizable by a zero-output residual block at location l:

$$\mathcal{U}_l = \left\{ u_l : u_l(z_l) = \frac{\partial}{\partial \eta} h_{l,\eta}(z_l)\Big|_{\eta=0} \right\}.$$

We assume that each residual tangent family is closed under nonnegative scalar multiplication:

$$u \in \mathcal{U}_l, \quad a \geq 0 \quad \Longrightarrow \quad au \in \mathcal{U}_l.$$

This property holds when the tangent family is generated by differentiable parameter paths whose initial velocities may be rescaled. Define the normalized first-order improvement capacity at layer l by

$$\Gamma_l = \sup_{\substack{u_l \in \mathcal{U}_l \\ \|u_l\|_{L^2(\mathcal{D})} \leq 1}} -\mathbb{E}\left[q_l^\top u_l\right].$$

A positive value $\Gamma_l > 0$ means that the residual module can realize a strict population-risk descent direction at that location.

Consider an early location l and a later location $r > l$. Let

$$\Phi_{l \to r} = T_{r-1} \circ \cdots \circ T_l$$

be the downstream mapping from z_l to z_r, and define

$$A_{l \to r}(x) = J_{\Phi_{l \to r}}(z_l(x)).$$

By the chain rule,

$$q_l(x, y) = A_{l \to r}(x)^\top q_r(x, y).$$

The following proposition formalizes when an early residual block is at least as capable of producing a first-order improvement as a later block.

Proposition 6 (Early-layer dominance under residual-tangent coverage)**.** *Let $l < r$. Suppose that there exists $\kappa_{l,r} > 0$ such that, for every $u_r \in \mathcal{U}_r$, there exists $u_l \in \mathcal{U}_l$ satisfying*

$$A_{l \to r}(x) u_l(z_l(x)) = u_r(z_r(x))$$

almost surely, and

$$\|u_l\|_{L^2(\mathcal{D})} \leq \kappa_{l,r}^{-1} \|u_r\|_{L^2(\mathcal{D})}.$$

Then

$$\Gamma_l \geq \kappa_{l,r} \Gamma_r.$$

In particular, if the later location admits a strict first-order descent direction, then the earlier location also admits a strict first-order descent direction.

Proof. Take any $u_r \in \mathcal{U}_r$ satisfying

$$\|u_r\|_{L^2(\mathcal{D})} \leq 1.$$

By the residual-tangent coverage assumption, there exists $u_l \in \mathcal{U}_l$ such that

$$A_{l \to r} u_l = u_r, \qquad \|u_l\|_{L^2(\mathcal{D})} \leq \kappa_{l,r}^{-1}.$$

Define

$$\widetilde{u}_l = \kappa_{l,r} u_l.$$

The nonnegative scaling closure ensures that $\widetilde{u}_l \in \mathcal{U}_l$. Moreover, $\|\widetilde{u}_l\|_{L^2(\mathcal{D})} \leq 1$, and the chain-rule identity gives

$$-\mathbb{E}\left[q_l^\top \widetilde{u}_l\right] = -\kappa_{l,r} \mathbb{E}\left[q_r^\top A_{l \to r} u_l\right].$$

Since $A_{l \to r} u_l = u_r$, it follows that

$$-\mathbb{E}\left[q_l^\top \widetilde{u}_l\right] = -\kappa_{l,r} \mathbb{E}\left[q_r^\top u_r\right].$$

Taking the supremum over all admissible u_r yields

$$\Gamma_l \geq \kappa_{l,r} \Gamma_r.$$

Therefore, any first-order descent direction realizable at the later location can also be realized at the earlier location, up to the conditioning factor $\kappa_{l,r}$. □

Proposition 6 provides the main representational reason for selecting an early ResNet block. Early ResNet representations retain a high spatial resolution and a comparatively large number of local degrees of freedom. A standard BasicBlock residual branch can be expressed as

$$h_{U,V}(z) = \mathrm{Conv}_{\mathrm{out}}\left(\psi_U(z)\right),$$

which is an instance of the abstract realizability form

$$h_{U,V}(z) = V\psi_U(z).$$

The feature map $\psi_U(z)$ contains the preceding convolutions, normalization operations, and nonlinearities, while the final convolutional projection plays the role of V. At an early stage, this residual branch acts on a high-resolution feature map and therefore generates a broad family of spatially structured tangent directions.

As the representation passes through later stages, stride-two downsampling and channel transformations compress and reorganize the spatial degrees of freedom. Under the tangent-coverage condition in Proposition 6, the image of the early residual tangent family under the downstream Jacobian contains the tangent family available at the later location:

$$\mathcal{U}_r \subseteq A_{l\to r}\mathcal{U}_l.$$

Thus, the early module can reproduce any perturbation available to the later module after downstream propagation, while also retaining additional perturbation directions that may be removed by spatial downsampling or later feature compression. This establishes a precise sense in which the early residual location is representationally at least as expressive for first-order residual expansion.

The finite-sample alignment geometry gives a second justification. At layer l, define

$$\bar{\mu}_l = \mathbb{E}[q_l], \qquad \Sigma_l = \mathrm{Cov}(q_l),$$

and

$$d_{\parallel}^{(l)} = \frac{\|\bar{\mu}_l\|_2^4}{\bar{\mu}_l^\top \Sigma_l \bar{\mu}_l}, \qquad d_2^{(l)} = \frac{\|\bar{\mu}_l\|_2^4}{\mathrm{tr}(\Sigma_l^2)}.$$

Suppose that the candidate ResNet locations satisfy the uniform covariance controls

$$\lambda_{\max}(\Sigma_l) \le \Lambda, \qquad \mathrm{tr}(\Sigma_l^2) \le T,$$

where Λ and T do not increase across the compared locations. Then

$$\bar{\mu}_l^\top \Sigma_l \bar{\mu}_l \le \Lambda \|\bar{\mu}_l\|_2^2,$$

and therefore

$$d_{\parallel}^{(l)} \ge \frac{\|\bar{\mu}_l\|_2^2}{\Lambda}.$$

Similarly,

$$d_2^{(l)} \ge \frac{\|\bar{\mu}_l\|_2^4}{T}.$$

Consequently,

$$d_{\mathrm{align}}^{(l)} \ge \min\left\{\frac{\|\bar{\mu}_l\|_2^2}{\Lambda}, \frac{\|\bar{\mu}_l\|_2^4}{T}\right\}.$$

The original residual-expansion analysis predicts that the remaining first-order signal becomes weaker as the model approaches the deepest-model regime. If an early candidate location l and a later location r satisfy

$$\|\bar{\mu}_l\|_2 \ge \|\bar{\mu}_r\|_2,$$

then the preceding lower bound gives

$$\min\left\{\frac{\|\bar{\mu}_l\|_2^2}{\Lambda}, \frac{\|\bar{\mu}_l\|_2^4}{T}\right\} \ge \min\left\{\frac{\|\bar{\mu}_r\|_2^2}{\Lambda}, \frac{\|\bar{\mu}_r\|_2^4}{T}\right\}.$$

Thus, under comparable covariance control, the early layer has a no-weaker guaranteed lower certificate for the effective alignment dimension. The corresponding finite-sample misalignment guarantee satisfies

$$\Pr\left(\mu_M^{(l)\top} g_K^{(l)} \le 0\right) \le \frac{1}{1 + n_{\mathrm{eff}} d_{\mathrm{align}}^{(l)}},$$

so a stronger early-layer signal leads to a tighter guaranteed alignment certificate.

The representational and statistical arguments are complementary. The tangent-coverage result shows that an early residual block can reproduce later first-order perturbations through the remaining network, while the signal-to-noise argument shows that an early location is preferable when the population first-order signal has not yet entered the weak-signal regime. This is consistent with the deepest-model mechanism of the original analysis, according to which the remaining activation-gradient signal becomes progressively weaker as residual depth is exhausted.

For ResNet-20, we therefore measure the activation gradient at the output of the first residual block. This location retains the full spatial resolution of the first residual stage, provides a rich realizable residual tangent family, and leaves the subsequent residual stages available to transform the inserted perturbation into a prediction-level change. The choice also follows the early-block regime used in the original ResNet covariance experiment. It should not be interpreted as an unconditional theorem that the first residual block maximizes d_{align} for every trained ResNet. Rather, it is the architecture-specific location favored by residual-tangent coverage, non-exhausted first-order signal, and controlled within-family width comparison.

M.4 Layer-wise Component Decomposition of the Effective Alignment Dimension

The component-wise decomposition $d_{\|}$, d_2, and $d_{\mathrm{align}} = \min\{d_{\|}, d_2\}$ is reported at the final residual representation, aggregated per width, in Table 8 of this appendix (and, for Pythia, in Table 9). The present subsection is purely descriptive: it provides additional depth-resolved observations and data points by reporting the same two relative-noise components at each of the recorded blocks (a representative subset of the 32-block stack), for all six residual-stream widths $N \in \{512, 1024, 1536, 2048, 2560, 3072\}$, so that the reader can see how $d_{\|}$ and d_2 evolve with depth. Each entry is the mean $\pm$ standard deviation obtained by first aggregating over independently sampled evaluation subsets within each trained model and then summarising across the eight independent training seeds.

Scope. This subsection is descriptive and carries no proof obligation; we recall two facts established elsewhere in this appendix only to prevent misreading, not to re-derive them. The measurement location is fixed by the minimal-transport principle, and the layer-wise theory is existential over admissible insertion layers (see the section *Architecture-Specific Selection of Activation-Gradient Measurement Locations*, in particular its paragraphs *Why the Final Residual Representation Is Measured in LLMs* and *Existential nature of the layer-wise theory*, Equations 59–61; the main text states only the resulting measurement convention, without deriving it). Consequently, the values below are transport-confounded measurements at intermediate depths, and—because the existence statement requires only a single admissible layer to satisfy the certificate, not every layer—the small shallow-layer values of d_2 and the mid-to-late regressions of $d_{\|}$ are not counterexamples to it: they merely record where the separately reported relative-noise components are less favourable under the present sampling budget. This is exactly the layer-dependent Jacobian-transport effect analysed in that section, and the component tables neither re-justify the final-block choice nor require the alignment condition to hold at every block.

Reporting convention. For each (layer, width) cell, we report separately aggregated means and standard deviations for $d_{\|}$ and d_2. We do *not* compute d_{align} by taking the minimum of these two aggregate means. In general,

$$\overline{\min\{d_{\|}, d_2\}} \neq \min\{\overline{d_{\|}}, \overline{d_2}\},$$

and the multistage aggregation over evaluation subsets and training seeds makes this distinction essential. Therefore, even though the reported mean of d_2 is smaller than the reported mean of $d_{\|}$ in every cell, Table 10 reports only d_2, not d_{align}. The d_{align} values in Table 8 are computed at the evaluation-unit level before aggregation and cannot be recovered from the two layer-wise component tables.

Depth and width trends (descriptive). The covariance-energy component d_2 grows overall with both depth and width, rising from roughly 0.5–1.9 at layer 1 to roughly 3.7–8.9 at layer 32, with the width ordering preserved at every layer and the width-induced gain amplified in the deeper blocks

(at layer 32 the gap between $N = 512$ and $N = 3072$ exceeds five units, versus about 1.4 units at layer 1); it is not strictly monotone in depth, exhibiting local dips in the intermediate blocks for the middle widths (e.g. $N \in \{1536, 2048, 2560\}$ between layers 3 and 17). The directional component $d_{\|}$ starts substantially larger in the shallow blocks (about 6.7–9.1) and increases more slowly with depth, and likewise exhibits local regressions in the mid-to-late blocks for the intermediate widths (e.g. layers 30–32 at $N \in \{1024, 1536, 2048, 2560\}$). As noted above, both the shallow-layer behaviour and these local non-monotonicities reflect the anisotropic, sample-dependent Jacobian transport accumulated by the downstream blocks; the resulting layer-wise quantities are therefore transport-confounded descriptive measurements rather than de-confounded certificate inputs. Read in this descriptive sense, the tables simply document how the two relative-noise components vary across the stack and confirm that the width-induced increase seen at the final block is not confined to that single location.

	Residual-stream width N					
Layer	512	1024	1536	2048	2560	3072
1	0.518 ± 0.052	0.692 ± 0.061	0.882 ± 0.078	1.170 ± 0.095	1.554 ± 0.112	1.929 ± 0.146
2	0.622 ± 0.060	0.791 ± 0.070	1.078 ± 0.088	1.365 ± 0.105	1.865 ± 0.125	2.234 ± 0.169
3	0.829 ± 0.075	0.989 ± 0.085	1.274 ± 0.102	1.559 ± 0.120	2.072 ± 0.142	2.539 ± 0.192
7	1.140 ± 0.095	1.384 ± 0.110	1.176 ± 0.135	1.052 ± 0.125	1.120 ± 0.130	2.843 ± 0.215
8	1.244 ± 0.102	1.483 ± 0.118	1.274 ± 0.142	1.145 ± 0.132	1.215 ± 0.138	3.046 ± 0.231
11	1.555 ± 0.120	1.780 ± 0.135	1.568 ± 0.155	1.420 ± 0.145	1.490 ± 0.150	3.859 ± 0.292
14	1.866 ± 0.140	2.076 ± 0.155	2.646 ± 0.185	3.021 ± 0.205	3.833 ± 0.235	4.874 ± 0.369
15	1.970 ± 0.148	2.175 ± 0.162	2.743 ± 0.192	3.119 ± 0.212	3.937 ± 0.242	5.077 ± 0.384
17	2.073 ± 0.155	2.373 ± 0.172	2.058 ± 0.195	1.905 ± 0.185	1.980 ± 0.190	4.570 ± 0.346
20	2.384 ± 0.175	2.571 ± 0.190	3.135 ± 0.220	3.509 ± 0.240	4.351 ± 0.270	5.585 ± 0.423
25	2.903 ± 0.210	3.065 ± 0.225	3.723 ± 0.255	4.094 ± 0.275	5.076 ± 0.305	6.601 ± 0.500
30	3.421 ± 0.245	3.559 ± 0.260	4.409 ± 0.290	4.776 ± 0.310	5.802 ± 0.340	8.022 ± 0.607
31	3.629 ± 0.260	3.856 ± 0.275	4.801 ± 0.305	5.166 ± 0.325	6.216 ± 0.355	8.530 ± 0.646
32	3.700 ± 0.383	4.198 ± 0.316	5.283 ± 1.093	5.698 ± 0.478	6.420 ± 1.324	8.886 ± 0.611

Table 10: Layer-wise covariance-energy component d_2 across Transformer blocks and residual-stream widths. Entries are mean $\pm$ standard deviation across eight trained models. These depth-resolved component values are descriptive and transport-confounded; they are not reported as d_{align}.

	Residual-stream width N					
Layer	512	1024	1536	2048	2560	3072
1	6.739 ± 0.329	6.921 ± 0.350	7.349 ± 0.410	7.797 ± 0.450	8.806 ± 0.520	9.139 ± 0.610
2	7.257 ± 0.355	7.415 ± 0.375	7.839 ± 0.435	8.285 ± 0.475	9.324 ± 0.545	9.647 ± 0.635
3	7.775 ± 0.380	7.910 ± 0.400	8.328 ± 0.460	8.772 ± 0.500	9.842 ± 0.570	10.155 ± 0.660
7	8.294 ± 0.405	8.206 ± 0.425	7.741 ± 0.485	7.520 ± 0.465	7.850 ± 0.475	9.342 ± 0.580
8	8.501 ± 0.416	8.404 ± 0.436	7.937 ± 0.496	7.710 ± 0.476	8.040 ± 0.486	9.647 ± 0.590
11	8.812 ± 0.431	8.701 ± 0.451	8.230 ± 0.511	8.005 ± 0.491	8.330 ± 0.501	9.952 ± 0.605
14	8.916 ± 0.436	8.898 ± 0.456	9.210 ± 0.516	9.552 ± 0.536	10.567 ± 0.606	10.764 ± 0.695
15	9.123 ± 0.446	9.096 ± 0.466	9.406 ± 0.526	9.746 ± 0.546	10.775 ± 0.616	10.967 ± 0.705
17	9.538 ± 0.466	9.393 ± 0.486	8.916 ± 0.546	8.690 ± 0.526	9.020 ± 0.536	10.663 ± 0.645
20	9.330 ± 0.456	9.294 ± 0.476	9.602 ± 0.536	9.941 ± 0.556	10.982 ± 0.626	11.170 ± 0.715
25	9.849 ± 0.481	9.689 ± 0.501	9.994 ± 0.561	10.234 ± 0.581	11.189 ± 0.651	11.373 ± 0.740
30	10.575 ± 0.517	9.096 ± 0.461	9.602 ± 0.521	8.187 ± 0.401	8.392 ± 0.411	10.459 ± 0.630
31	10.886 ± 0.532	9.393 ± 0.476	9.896 ± 0.536	8.479 ± 0.416	8.703 ± 0.426	10.764 ± 0.645
32	10.797 ± 0.832	9.744 ± 0.645	10.373 ± 0.568	8.908 ± 0.240	8.638 ± 0.305	10.876 ± 0.442

Table 11: Layer-wise directional component $d_{\|}$ across Transformer blocks and residual-stream widths, using the same aggregation as Table 10. The directional component exceeds d_2 in every reported cell; the depth-resolved values are descriptive and transport-confounded.