NORMALIZATION AND THE SCALAR STRUCTURE OF DROPOUT

Anonymous authors
Paper under double-blind review

ABSTRACT

Dropout randomly changes individual features in a neural network. We ask when its effect after normalization can be approximated by one common scalar rescaling. For a trained network and a given input, normalization divides the features by a shared factor. When many small independent contributions determine this factor, its relative variation can become small. We replace it by a value computed directly from the features and dropout probability. At this fixed radius, the scalar minimizes mean squared error. It rescales the part of the output affected by normalization and leaves the fixed part unchanged. We then quantify how much of the change caused by dropout this scalar removes in mean square and how much random variation remains. Under replication that preserves predictions and uses independent copy masks, the output error is bounded in probability at the inverse square root of the number of copies. For RMS normalization, a nonzero limiting covariance rules out a faster rate. Experiments on MNIST networks, a pretrained FaceNet embedding, and ten language models test the theory. In the language models, scalar rescaling reduces error against Monte Carlo means of probability vectors. In FaceNet, it reduces error against Monte Carlo means of cosine scores. Individual dropout outputs can still remain highly variable.

1 INTRODUCTION

When dropout is sampled at inference time, the same trained network is evaluated under different random masks and produces a distribution of outputs. Monte Carlo estimation of its mean requires repeated forward passes (Gal & Ghahramani, 2016). We ask when one scalar, computed without sampled masks, suffices to approximate individual dropout outputs or their mean.

LayerNorm and RMSNorm suggest a candidate scalar through their shared denominator that depends on the mask (Ba et al., 2016; Zhang & Sennrich, 2019). Through a fixed linear head, this denominator scales the part of every output that depends on the representation. Dropout also perturbs the numerator and can change the relative outputs. The approximation error is small when the common coefficient varies little and the numerator change reaching the output is small.

Our objective is a theoretical description of dropout after normalization. For a fixed trained network and input, only the mask is random. We derive the scalar from stored features and the dropout probability, without labels or sampled masks, and quantify its approximation error and relative reduction in mean squared error.

High dimension helps through averaging: many independent mask coordinates contributing at comparable scales can stabilize the normalization factor (Vershynin, 2018; Wainwright, 2019). Shared masks or a dominant coordinate can prevent its relative fluctuations from decreasing, so dimension alone is insufficient.

Averaging independent dropout outputs reduces variability around their mean, not the discrepancy between that mean and the analytic target. A scalar can therefore approximate the mean accurately while individual outputs remain highly variable.

Contributions. First, we derive an exact radius-numerator decomposition, finite-dimensional concentration bounds, and a law relating the mean shift to dispersion. At fixed radius, the analytic coefficient uniquely minimizes MSE when the feature contribution is nonzero, and the law identifies

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.

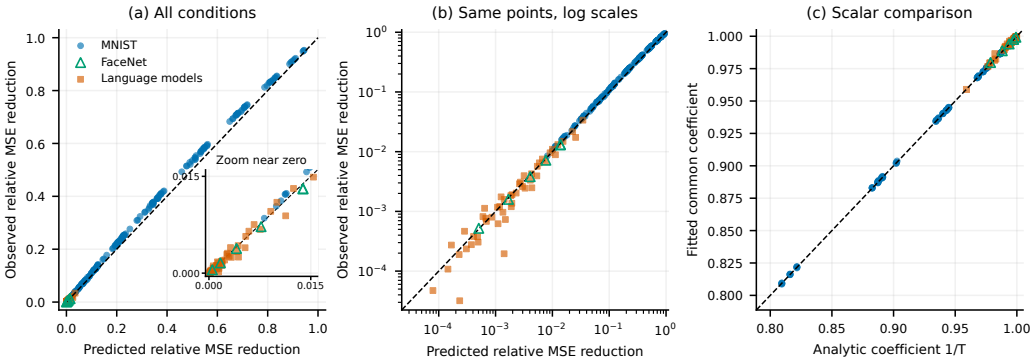


Figure 1: **Predicted MSE reduction and optimal scalar.** Panels (a) and (b) compare predictions at fixed radius with relative MSE reductions from actual outputs, on linear and logarithmic axes. Each point is one native MNIST or independent-copy condition, one FaceNet condition, or one language model condition. Panel (c) compares analytic and fitted common coefficients. The diagonal marks exact agreement. Detailed comparisons appear in Appendices F.1 and F.2.

the irreducible residual. Complete proofs appear in Appendix B, with a ReLU extension under explicit conditions in Appendix D.

Second, replication preserves the original prediction and isolates the effect of mask independence. With independent masks, the approximation error is of order at most the inverse square root of the number of copies. Under RMS normalization and a nonzero limiting covariance, a Gaussian limit rules out a faster rate. Shared masks preserve the fluctuations. Changes of coordinates show that dropout also depends on how the perturbed features and the output map are expressed in the chosen basis.

Finally, experiments test quantitative predictions and their limits on 27 MNIST networks, FaceNet on full LFW (Schroff et al., 2015; Huang et al., 2007), and ten GPT-2, OPT, T5, and BART checkpoints (Radford et al., 2019; Zhang et al., 2022; Raffel et al., 2020; Lewis et al., 2020). They assess approximation error in trained networks, distinguishing mean outputs from individual dropout outputs. We evaluate on new masks and tokens, and test scalar approximation with dropout active throughout the network.

2 Related work

Dropout as regularization and averaging. Dropout began as a regularizer and an implicit average over random models (Srivastava et al., 2014). Theory connected it to adaptive regularization and efficient averaging in linear and shallow models (Baldi & Sadowski, 2013; Wager et al., 2013; Wang & Manning, 2013). Later work studied equal influence, nuclear-norm effects, optimization, capacity, and feature interactions (Mianjy et al., 2018; Mianjy & Arora, 2019; 2020; Wei et al., 2020; Arora et al., 2021; Lengerich et al., 2022).

Stochastic prediction and related masks. Variational, Monte Carlo, and Concrete dropout connect random masks to probabilistic prediction or learned mask probabilities (Kingma et al., 2015; Gal & Ghahramani, 2016; Gal et al., 2017). R-Drop compares two random passes (Liang et al., 2021). DropConnect, DropBlock, and stochastic depth instead act on weights, spatial regions, or complete residual paths (Wan et al., 2013; Ghiasi et al., 2018; Huang et al., 2016). Our replication keeps the function fixed as in Net2Net (Chen et al., 2016) and changes only the dependence between copied masks.

Normalization and angular outputs. Dropout placement around BatchNorm and residual blocks can affect training (Li et al., 2019; Kim et al., 2023). We study dropout followed by LayerNorm, RMSNorm, or Euclidean normalization (Ba et al., 2016; Zhang & Sennrich, 2019), deriving an analytic scalar from the normalization radius. FaceNet, CosFace, and ArcFace provide the angular

setting of normalized embeddings and cosine scores (Schroff et al., 2015; Wang et al., 2018; Deng et al., 2019).

High-dimensional probability and concentration of measure. High-dimensional probability studies when aggregate quantities have small relative fluctuations (Vershynin, 2018; Wainwright, 2019). The Hanson–Wright inequality controls quadratic forms in independent sub-Gaussian variables through Frobenius and operator norms (Rudelson & Vershynin, 2013). We use it to control the normalization radius before separating the common scalar shift from residual output variation in the model with a fixed radius.

Relation to temperature scaling. Temperature scaling for calibration fits a positive rescaling of logits on labeled validation data (Guo et al., 2017). Our coefficient is computed per input from the representation and mask variance. When the fixed logit component vanishes, the transformations have the same algebraic form.

3 From coordinate dropout to an analytic scalar

Local setting. Fix a trained network, one input, and one dropout location immediately followed by normalization and a linear head. Let p be the feature dimension and m the output dimension. The outputs are logits, except in the FaceNet experiment where they are cosine scores. Weights and representations without dropout are fixed. Only the dropout mask is random, and all probabilities and expectations below are taken over this mask. Appendix A gives a complete notation guide.

Write the representation before normalization as $h = x + u$, where $h, x, u \in \mathbb{R}^p$. Dropout is applied to u, while x is an additive bypass contribution. Direct dropout on the complete representation is the case $x = 0$ and $u = h$. Writing the inverted mask as $1 + \zeta_i$ gives

$$h_\zeta = x + (\mathbf{1}_p + \zeta) \odot u = x + u + D\zeta = h + D\zeta, \qquad D \coloneqq \operatorname{diag}(u) \in \mathbb{R}^{p\times p}, \tag{1}$$

where $D\zeta$ is the random change. For inverted Bernoulli dropout,

$$\zeta_i \coloneqq \beta_i/q - 1, \qquad \beta_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(q), \qquad \rho \coloneqq (1-q)/q, \tag{2}$$

where $q \in \,]0, 1[$ is the keep probability. The factor $1/q$ preserves the mean, so $\mathbb{E}\zeta = 0$ and $\mathbb{E}(\zeta\zeta^\top) = \rho I_p$. The endpoint $q = 1$ has no dropout. Results requiring concentration also cover independent centered masks with common variance ρ under the sub-Gaussian condition stated below.

Normalization and logits. Normalization receives h without dropout and $h_\zeta = h + D\zeta$ with dropout. RMSNorm uses one scalar denominator. LayerNorm first centers the coordinates, as encoded by the projector $P = P^\top = P^2$:

$$P = I_p \quad \text{for RMS or Euclidean normalization,} \qquad P = I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p^\top \quad \text{for LayerNorm.} \tag{3}$$

We collect all fixed linear factors in $A \in \mathbb{R}^{m\times p}$ with $AP = A$, and all fixed shifts in $c \in \mathbb{R}^m$. In this Euclidean radius convention, with stabilizer $\varepsilon > 0$, the two cases are

$$\begin{aligned} \text{without dropout:} \quad & r_0 \coloneqq \sqrt{\|Ph\|_2^2 + \varepsilon}, \quad \ell_0 \coloneqq c + \frac{APh}{r_0}, \\ \text{with dropout:} \quad & r_\zeta \coloneqq \sqrt{\|Ph_\zeta\|_2^2 + \varepsilon}, \quad \ell_\zeta \coloneqq c + \frac{APh_\zeta}{r_\zeta}. \end{aligned} \tag{4}$$

The scalars r_0 and r_ζ are the regularized lengths of Ph and Ph_ζ, hence the term normalization radii. Each is shared by all normalized coordinates, so changing only the denominator would multiply $\ell_0 - c$ by r_0/r_ζ. The perturbation also enters the numerator, as the exact decomposition below shows.

For softmax, ℓ_0 and ℓ_ζ are centered logits, and centering preserves probabilities. For FaceNet, they are cosine scores without centering. Appendix C gives the exact implementation correspondence.

Scalar approximation objective. The output without dropout has a fixed part c and a part produced by the normalized representation. We seek one scalar a, independent of the drawn mask, to rescale only the second part:

$$\underbrace{\ell_\zeta - c}_{\text{with dropout}} \approx a \underbrace{(\ell_0 - c)}_{\text{without dropout}}.$$

To identify a candidate scalar, separate the unchanged numerator from the dropout perturbation in (4):

Proposition 1 (Exact radius-numerator decomposition)**.**

$$\ell_\zeta = c + \underbrace{\frac{r_0}{r_\zeta}(\ell_0 - c)}_{\textit{scalar rescaling}} + \underbrace{\frac{APD\zeta}{r_\zeta}}_{\textit{numerator effect}} . \tag{5}$$

The ratio r_0/r_ζ is shared by all coordinates of $\ell_0 - c$, while $APD\zeta/r_\zeta$ is the normalized numerator perturbation. A fixed scalar can work when the ratio varies little and this perturbation is small. To choose a coefficient before drawing a mask, we compute the mean squared radius. Expanding $r_\zeta^2 = \|P(h + D\zeta)\|_2^2 + \varepsilon$ separates the linear and quadratic mask contributions:

$$r_\zeta^2 = \underbrace{\|Ph\|_2^2 + \varepsilon}_{r_0^2} + \underbrace{2(DPh)^\top \zeta}_{\substack{\text{centered} \\ \text{fluctuation}}} + \underbrace{\zeta^\top DPD\zeta}_{\substack{\text{quadratic} \\ \text{mask term}}} . \tag{6}$$

The middle term has mean zero. Since $\mathbb{E}(\zeta\zeta^\top) = \rho I_p$, the last term has mean $\rho \operatorname{Tr}(DPD)$. Therefore

$$\bar{r}^2 \coloneqq \mathbb{E} r_\zeta^2 = r_0^2 + \rho \operatorname{Tr}(DPD), \qquad T \coloneqq \frac{\bar{r}}{r_0} = \sqrt{1 + \frac{\rho \operatorname{Tr}(DPD)}{r_0^2}} \geq 1. \tag{7}$$

Replacing r_0/r_ζ by $1/T = r_0/\bar{r}$ and omitting the random numerator perturbation defines the scalar target. Subtracting it from the exact output separates the two approximation errors:

$$\ell_T \coloneqq c + \frac{APh}{\bar{r}} = c + \frac{\ell_0 - c}{T}, \tag{8}$$

$$\ell_\zeta - \ell_T = \underbrace{\left(\frac{r_0}{r_\zeta} - \frac{1}{T}\right)(\ell_0 - c)}_{\text{error from fixing the radius}} + \underbrace{\frac{APD\zeta}{r_\zeta}}_{\text{numerator error}} . \tag{9}$$

The target rescales only the component that depends on the features, reducing to ℓ_0/T when $c = 0$. For a fixed input, network, and dropout probability, ℓ_T is deterministic, whereas the actual output retains the random numerator and radius.

For RMSNorm, $P = I_p$, so the expression for T^2 simplifies to

$$T^2 = 1 + \rho \frac{\|u\|_2^2}{\|x + u\|_2^2 + \varepsilon}. \tag{10}$$

Both $h = x + u$ and u are stored during one forward pass without sampled dropout. Hence T requires no labels and no sampled masks.

4 Error bounds for the scalar approximation

Equation (9) leaves two tasks: controlling the common coefficient through the radius and the numerator error through $APD\zeta$. These errors need not be orthogonal, even after expectation. Theorem 4 separates mean shift and dispersion orthogonally after fixing the radius.

Concentration of the common coefficient. To control r_0/r_ζ, we first control r_ζ^2 around $\bar{r}^2$. Squaring removes the square root and exposes sums to which concentration applies. For $\rho > 0$, standardize the mask as $\xi \coloneqq \zeta/\sqrt{\rho}$. Inverted Bernoulli dropout gives independent, centered, unit-variance coordinates. The bounds below only require these properties and the tail condition

$$\|\xi_i\|_{\psi_2} \leq K, \qquad \|X\|_{\psi_2} \coloneqq \inf\{s > 0 : \mathbb{E} e^{X^2/s^2} \leq 2\}. \tag{11}$$

This is the sub-Gaussian norm (Vershynin, 2018). The condition gives exponential tails. It holds for Gaussian noise and, for fixed q, for inverted Bernoulli dropout with a bound $K(q)$ independent of p.

With the deterministic symmetric matrix $B \coloneqq DPD$, the exact centered expansion is

$$r_\zeta^2 - \bar{r}^2 = 2\sqrt{\rho}\,\langle DPh, \xi\rangle + \rho\{\xi^\top B\xi - \operatorname{Tr}(B)\}. \tag{12}$$

The first term is a linear form. The second is a centered quadratic form. They are controlled respectively by a sub-Gaussian tail bound and the Hanson–Wright inequality (Rudelson & Vershynin, 2013, Theorem 1.1). For deterministic a, B, and G, and $t \in [1, +\infty[$, we use

$$\mathbb{P}\big(|\langle a, \xi\rangle| > CK\|a\|_2\sqrt{t}\big) \le 2e^{-t}, \tag{13}$$

$$\mathbb{P}\big(|\xi^\top B\xi - \mathrm{Tr}(B)| > CK^2\{\|B\|_{\mathrm{F}}\sqrt{t} + \|B\|_{\mathrm{op}}t\}\big) \le 2e^{-t}. \tag{14}$$

$$\mathbb{P}\big(\|G\xi\|_2 > CK^2\{\|G\|_{\mathrm{F}} + \|G\|_{\mathrm{op}}\sqrt{t}\}\big) \le 2e^{-t}.$$

The constant C is absolute (Vershynin, 2018). With $a = DPh$ and $B = DPD$, the first two bounds control $r_\zeta^2 - \bar{r}^2$. The Frobenius norm combines directions, while the operator norm controls the largest directional contribution. Dividing the bound by $\bar{r}^2$ controls relative fluctuations of the squared radius.

Visible numerator variation. The matrix APD maps mask perturbations to the output numerator. Its squared Frobenius norm sums squared coordinate responses, while its operator norm measures the largest response to a unit Euclidean perturbation. Since $\mathbb{E}(\zeta\zeta^\top) = \rho I_p$, the mean squared numerator change is $\rho\|APD\|_{\mathrm{F}}^2$. To compare this change with the numerator without dropout and assess whether a few directions dominate, define for $APh \neq 0$

$$\mathcal{V} := \frac{\mathbb{E}\|APD\zeta\|_2^2}{\|APh\|_2^2} = \frac{\rho\|APD\|_{\mathrm{F}}^2}{\|APh\|_2^2}, \qquad r_{\mathrm{vis}} := \frac{\|APD\|_{\mathrm{F}}^2}{\|APD\|_{\mathrm{op}}^2}. \tag{15}$$

When $APD \neq 0$, $1/r_{\mathrm{vis}}$ is the largest directional variance of the output numerator divided by its total variance. A large r_{vis} therefore prevents any single direction from carrying a large share of that variance. Set $r_{\mathrm{vis}} = +\infty$ when $APD = 0$. The relative comparison is undefined when $APh = 0$, because the feature contribution without dropout vanishes.

Apply the third bound with $G = APD$. A union bound makes all three estimates simultaneous with probability at least $1 - 6e^{-t}$. For $t \in [1, +\infty[$, collect them in

$$\mathfrak{R}_t := C\left[K\sqrt{\rho}\,\|DPh\|_2\sqrt{t} + K^2\rho\left(\|DPD\|_{\mathrm{F}}\sqrt{t} + \|DPD\|_{\mathrm{op}}t\right)\right], \tag{16}$$

$$\mathfrak{D}_t := CK^2\sqrt{\rho}\left(\|APD\|_{\mathrm{F}} + \|APD\|_{\mathrm{op}}\sqrt{t}\right). \tag{17}$$

To assess whether radius fluctuations are small compared with the reference squared radius, define $\chi_t := \mathfrak{R}_t/\bar{r}^2$. The concentration estimates control r_ζ^2, while the normalized output contains $1/r_\zeta$. Theorem 2 turns the bound on the squared radius into a bound on its inverse and combines it with the numerator bound $\mathfrak{D}_t$.

Theorem 2 (Finite-dimensional probability bound)**.** *For $t \in [1, +\infty[$, under the assumptions above, with probability at least $1 - 6e^{-t}$,*

$$|r_\zeta^2 - \bar{r}^2| \le \mathfrak{R}_t, \qquad \|APD\zeta\|_2 \le \mathfrak{D}_t. \tag{18}$$

If $\chi_t < 1$, then

$$\left|\frac{r_0}{r_\zeta} - \frac{1}{T}\right| \le \frac{T^{-1}\chi_t}{\sqrt{1-\chi_t}(1+\sqrt{1-\chi_t})}, \tag{19}$$

$$\left\|\frac{APD\zeta}{r_\zeta}\right\|_2 \le \frac{\mathfrak{D}_t}{\sqrt{1-\chi_t}\,\bar{r}}. \tag{20}$$

If, in addition, $APh \neq 0$, then

$$\frac{\|\ell_\zeta - \ell_T\|_2}{\|(\ell_0 - c)/T\|_2} \le \frac{\chi_t}{\sqrt{1-\chi_t}(1+\sqrt{1-\chi_t})} + \frac{CK^2}{\sqrt{1-\chi_t}}\sqrt{\mathcal{V}}\left(1 + \sqrt{\frac{t}{r_{\mathrm{vis}}}}\right). \tag{21}$$

The bound on the inverse radius follows from $r_\zeta \ge \sqrt{1-\chi_t}\,\bar{r}$, which keeps the denominator away from zero, and the identity

$$\left|\frac{1}{r_\zeta} - \frac{1}{\bar{r}}\right| = \frac{|r_\zeta^2 - \bar{r}^2|}{r_\zeta\bar{r}(r_\zeta + \bar{r})}.$$

The same event also lower-bounds r_ζ in the error term from the numerator. The relative bound in (21) is obtained by dividing the absolute error by $\|(\ell_0 - c)/T\|_2 = \|APh\|_2/\bar{r}$, the size of the component that the scalar approximation reproduces from the features. The fixed vector c is reproduced exactly and is therefore excluded from this normalization. The two terms in (21) measure the radius replacement and the visible numerator change. The bound is small when χ_t and $K^2\sqrt{\mathcal{V}}(1+\sqrt{t/r_{\mathrm{vis}}})$ are small.

Radius stabilization with dimension. To see when the bound improves with p, compare $\bar{r}^2$, the mean squared radius, with its fluctuation. Suppose that $\|D\|_{\mathrm{op}} \le U$, $\|Ph\|_2 \le H\sqrt{p}$, and $\bar{r}^2 \ge c_0 p$. At fixed t, K, ρ, U, H, and c_0, the scaling and explicit bound are

$$\chi_t = \frac{\mathfrak{R}_t}{\bar{r}^2} = O(p^{-1/2}) \quad \text{since } \bar{r}^2 \ge c_0 p, \quad \mathfrak{R}_t = O(\sqrt{p}),$$
$$\chi_t \le \frac{C}{c_0}\left[K\sqrt{\rho}\,UH\sqrt{\frac{t}{p}} + K^2\rho U^2\left(\sqrt{\frac{t}{p}} + \frac{t}{p}\right)\right]. \tag{22}$$

Under these bounds, the mean squared radius grows at least linearly with p, whereas its fluctuation bound grows at most as $\sqrt{p}$. A dropout-affected coordinate whose scale grows with p can invalidate this separation, so dimension alone does not imply the rate.

The next corollary gives a uniform-tail sufficient condition along a sequence of models, for instance as the feature dimension increases.

Corollary 3 (A sufficient condition)**.** *Fix $\delta \in]0,1[$ and set $t = \log(6/\delta)$. Along any sequence of models satisfying the assumptions above with $A_n P_n h_n \neq 0$, if*

$$\sup_n K_n < \infty, \qquad \chi_{t,n} \longrightarrow 0, \qquad \mathcal{V}_n\left(1 + \frac{t}{r_{\mathrm{vis},n}}\right) \longrightarrow 0, \tag{23}$$

then, for every $\epsilon > 0$ and every sufficiently large n, the relative error in (21) is at most ϵ with probability at least $1-\delta$.

The bound on K_n controls the tails, $\chi_{t,n} \to 0$ stabilizes the radius, and the last condition removes the visible numerator perturbation. Together they give the stated scalar approximation.

Independent noise followed by a linear map. Appendix B extends the analysis to $h_\zeta = h + M\zeta$, as needed when dropout is followed by a linear map. This includes FaceNet with $M = W\,\mathrm{diag}(g)$, where the number of mask variables can differ from p.

5 Mean shift, dispersion, and the optimal scalar

The previous section bounds the distance between the random output ℓ_ζ and the scalar target ℓ_T. We now solve the scalar problem exactly at fixed radius, measure the resulting relative MSE reduction, and bound how both conclusions change when the random radius r_ζ is restored.

Model with a fixed radius. For ℓ_ζ, the numerator $APh + APD\zeta$ and the radius r_ζ depend on the same mask. Hence $\mathbb{E}\ell_\zeta$ cannot in general be obtained by averaging the numerator and inverse radius separately. The model with a fixed radius removes this dependence by keeping the complete random numerator and replacing only r_ζ with $\bar{r}$:

$$\widetilde{\ell}_\zeta = c + \frac{APh + APD\zeta}{\bar{r}} = \ell_T + \frac{APD\zeta}{\bar{r}}. \tag{24}$$

Since $\mathbb{E}\zeta = 0$, the mean of $\widetilde{\ell}_\zeta$ is exactly ℓ_T. Define the squared shift from ℓ_0 to this mean and the mean squared dispersion around it by

$$R_{\mathrm{shift}} := \|\ell_T - \ell_0\|_2^2 = \frac{(T-1)^2\|APh\|_2^2}{\bar{r}^2}, \qquad D_{\mathrm{vis}} := \mathbb{E}\|\widetilde{\ell}_\zeta - \ell_T\|_2^2 = \frac{\rho\|APD\|_{\mathrm{F}}^2}{\bar{r}^2}. \tag{25}$$

Mean square risk at fixed radius. We approximate $\widetilde{\ell}_\zeta$ by $c + a(\ell_0 - c)$, where the scalar a is independent of the drawn mask. Its mean squared error is $\mathcal{Q}_{\text{fr}}(a) := \mathbb{E}\|\widetilde{\ell}_\zeta - \{c + a(\ell_0 - c)\}\|_2^2$, where the expectation is over masks.

Theorem 4 (Optimal scalar at fixed radius)**.** *Assume $APh \neq 0$. For every $a \in \mathbb{R}$, the risk separates into a squared scalar mismatch from the mean ℓ_T and the dispersion around that mean:*

$$\mathcal{Q}_{\text{fr}}(a) = \left(a - T^{-1}\right)^2 \|\ell_0 - c\|_2^2 + D_{\text{vis}}, \tag{26}$$

The choice $a = 1$ uses ℓ_0 itself as the approximation. Its baseline risk is

$$\mathcal{Q}_{\text{fr}}(1) = \mathbb{E}\|\widetilde{\ell}_\zeta - \ell_0\|_2^2 = R_{\text{shift}} + D_{\text{vis}}, \tag{27}$$

Only the first term in (26) depends on a. It vanishes uniquely at $a = T^{-1}$, so

$$a_\star = T^{-1} \quad \textit{(unique)}, \qquad \min_a \mathcal{Q}_{\text{fr}}(a) = D_{\text{vis}}, \tag{28}$$

After optimization, the remaining MSE relative to the squared target size is

$$\mathcal{V} = \frac{\min_a \mathcal{Q}_{\text{fr}}(a)}{\|\ell_T - c\|_2^2} = \frac{D_{\text{vis}}}{\|\ell_T - c\|_2^2}. \tag{29}$$

This is the Pythagorean form of a bias-variance decomposition in output space, with no supervised target. The residual D_{vis} is irreducible only within the stated scalar family at fixed radius. The ratio $\mathcal{V}$ is dimensionless but need not lie in $[0, 1]$.

Relative reduction in mean squared error. To compare inputs or models, divide the MSE decrease by the baseline MSE. For a square-integrable random output L with $\mathbb{E}\|L - \ell_0\|_2^2 > 0$, define

$$\mathcal{E}(L) := \frac{\mathbb{E}\|L - \ell_0\|_2^2 - \mathbb{E}\|L - \ell_T\|_2^2}{\mathbb{E}\|L - \ell_0\|_2^2} = 1 - \frac{\mathbb{E}\|L - \ell_T\|_2^2}{\mathbb{E}\|L - \ell_0\|_2^2}. \tag{30}$$

The *relative* reduction is the MSE removed by ℓ_T divided by the baseline MSE. The values 1, 0, and a negative value mean complete removal, no improvement, and a larger error, respectively.

At fixed radius, Theorem 4 separates the baseline MSE into the removable mean shift R_{shift} and the residual dispersion D_{vis}. To compare their sizes, when $\mathcal{V} > 0$ define the dimensionless ratio $\Lambda := R_{\text{shift}}/D_{\text{vis}} = (T-1)^2/\mathcal{V}$.

Corollary 5 (Relative MSE reduction at fixed radius)**.** *If $\mathcal{V} > 0$, then*

$$\mathcal{E}(\widetilde{\ell}_\zeta) = \frac{R_{\text{shift}}}{R_{\text{shift}} + D_{\text{vis}}} = \frac{\Lambda}{1 + \Lambda}. \tag{31}$$

When $\Lambda \gg 1$, the scalar removes most of the MSE at fixed radius. When $\Lambda \ll 1$, the residual dispersion dominates. Both T and $\mathcal{V}$ are computed from fixed network quantities, without labels or a fitted scalar.

Effect of restoring the random radius. Theorem 4 and Corollary 5 are exact for $\widetilde{\ell}_\zeta$, not for the original output ℓ_ζ. To assess their accuracy for ℓ_ζ, define the mean squared distance between the two outputs by $\epsilon_{\text{fr}}^2 := \mathbb{E}\|\ell_\zeta - \widetilde{\ell}_\zeta\|_2^2$. When $R_{\text{shift}} + D_{\text{vis}} > 0$, normalize its square root as $\eta_{\text{fr}} := \epsilon_{\text{fr}}/\sqrt{R_{\text{shift}} + D_{\text{vis}}}$. The denominator is the root mean squared change at fixed radius from ℓ_0, so $\eta_{\text{fr}} \ll 1$ means that the error from fixing the radius is small on this scale. The next proposition controls scalar stability using ϵ_{fr} and relative-MSE stability using η_{fr}. For the original output, let a_{true} minimize $\mathcal{Q}(a) := \mathbb{E}\|\ell_\zeta - \{c + a(\ell_0 - c)\}\|_2^2$.

Proposition 6 (Stability for the original normalized output)**.** *If $APh \neq 0$, the scalar that minimizes MSE for the original output ℓ_ζ satisfies*

$$\left|a_{\text{true}} - T^{-1}\right| \leq \frac{\epsilon_{\text{fr}}}{\|\ell_0 - c\|_2}, \qquad \mathcal{Q}(T^{-1}) - \min_a \mathcal{Q}(a) \leq \epsilon_{\text{fr}}^2. \tag{32}$$

If, in addition, $\mathcal{V} > 0$ and $\eta_{\text{fr}} \leq 1/4$, then

$$\left|\mathcal{E}(\ell_\zeta) - \frac{\Lambda}{1 + \Lambda}\right| \leq 12\eta_{\text{fr}}. \tag{33}$$

The bounds in (32) control the difference between the optimal coefficients and the excess MSE from using $1/T$. Equation (33) controls the difference between the relative MSE reductions for the actual output and the output at fixed radius. Appendix B derives an explicit bound on ϵ_{fr} from Theorem 2.

6 HIGH-DIMENSIONAL STABILIZATION

To isolate the effect of mask independence, we increase the number of coordinates without changing the output without dropout. For integer $k \geq 1$ and $v \in \mathbb{R}^p$, replication gives k scaled copies per coordinate: $(L_k v)_{i,j} \coloneqq v_i/\sqrt{k}$, for $1 \leq i \leq p$ and $1 \leq j \leq k$. Set $h_k \coloneqq L_k h$, $u_k \coloneqq L_k u$, and $A_k \coloneqq AL_k^\top$, keeping c and the Euclidean stabilizer ε fixed. Shared masks also preserve every masked output, whereas independent copy masks average their contributions. For independent masks, assume the conditions preceding (11) with the same $\rho > 0$ and K for every k. Let $\mathcal{V}_k$ denote the ratio in (15) for this replicated model, with $\mathcal{V}_1$ its original value.

Theorem 7 (Averaging under independent replication)**.** *Under these independent copy masks, if $APh \neq 0$, then, for every fixed $t \in [1, +\infty[$,*

$$\mathcal{V}_k = \frac{\mathcal{V}_1}{k}, \qquad r_{\mathrm{vis},k} = r_{\mathrm{vis},1}, \tag{34}$$

$$\chi_{t,k} = O(k^{-1/2}), \qquad \|\ell_{k,\zeta} - \ell_{T_k}\|_2 = O_{\mathbb{P}}(k^{-1/2}). \tag{35}$$

Both rates hold for RMSNorm and LayerNorm, with constants independent of k. For RMSNorm, $T_k = T_1$. For LayerNorm, the centering correction vanishes, giving the limiting radius ratio $T_\infty \coloneqq \sqrt{1 + \rho\|u\|_2^2/r_0^2}$ and $T_k = T_\infty + O(k^{-1})$.

Each independent copy reaches the output numerator with amplitude $1/k$. The k squared amplitudes sum to $1/k$, giving $\mathcal{V}_k = \mathcal{V}_1/k$ and a numerator perturbation of typical size $O(k^{-1/2})$. For RMSNorm, if $T_1 > 1$ and $\mathcal{V}_1 > 0$, then $\Lambda_k = k\Lambda_1$ and

$$\mathcal{E}(\widetilde{\ell}_{k,\zeta}) = \frac{k\Lambda_1}{1 + k\Lambda_1} \longrightarrow 1. \tag{36}$$

Equation (36) is exact for the model with a fixed radius. Under these RMS assumptions and inverted Bernoulli dropout, Appendix B also proves that the relative MSE reduction for $\ell_{k,\zeta}$ converges to one. It separately gives finite-dimensional stability bounds, a Gaussian limit, and a bound for the mean softmax probabilities.

6.1 DEPENDENCE ON THE COORDINATE SYSTEM

Coordinate-wise dropout can depend on the basis even when the output without dropout is unchanged. Choose orthogonal Q, with $Q\mathbf{1}_p = \mathbf{1}_p$ for LayerNorm to preserve coordinate centering. Set $(h_Q, u_Q, A_Q) \coloneqq (Qh, Qu, AQ^\top)$ and $q_i \coloneqq Q^\top e_i$. These transformations preserve the output without dropout and T. The transformed visible-noise ratio, denoted by $\mathcal{V}(Q)$, compares the mean squared numerator perturbation with the squared numerator without dropout:

$$\mathcal{V}(Q) = \frac{\rho}{\|Ah\|_2^2} \sum_{i=1}^{p} (q_i^\top u)^2 \|Aq_i\|_2^2. \tag{37}$$

Term i pairs the squared component of u on axis i with the squared output change caused by that axis. The next theorem averages these products over a uniformly random basis.

Theorem 8 (Mean visibility over random coordinate systems)**.** *Assume $P = I_p$ and $Ah \neq 0$. For Haar-distributed $Q \in O(p)$,*

$$\mathbb{E}_Q \mathcal{V}(Q) = \frac{\rho}{\|Ah\|_2^2} \frac{\|u\|_2^2 \|A\|_{\mathrm{F}}^2 + 2\|Au\|_2^2}{p + 2}. \tag{38}$$

This is the exact mean over random bases, not a guarantee for every basis or for the complete normalized-output error. The factor $1/(p+2)$ alone is not a dimension rate because the other norms may depend on p.

A random basis generally spreads the direct output path over every axis. As a deterministic contrast, align the basis with the row and null spaces of A.

Proposition 9 (Output map supported on $\mathrm{rank}(A)$ axes)**.** *Assume $P = I_p$. If $r_A = \mathrm{rank}(A)$, there is an orthogonal Q such that $Aq_i = 0$ for every $i > r_A$. Thus only r_A axes directly affect the visible numerator. All axes may affect the radius.*

This support is minimal by rank, but need not maximize $\mathcal{V}(Q)$: $q_i^\top u$ also depends on the basis. Both results show that dropout perturbations can differ at fixed output without dropout and T. For centered softmax logits, $r_A \le m-1$, while language model output maps may use almost all feature directions. Further constructions, including rotations compatible with LayerNorm, appear in Appendix F.3.

7 Experiments in the two regimes

Analytic diagnostic. We test the predicted MSE reduction and approximation of Monte Carlo means. One deterministic pass stores h, u, and A before masks are drawn. Correlation measures association, MAE numerical discrepancy, and log–log slopes observed scaling (Figure 1). For protocols, estimators, and additional results, see Appendices E, E.1, and F.

Small MNIST models. For MNIST (LeCun et al., 1998), we test residual MLPs with RMSNorm, small vision transformers with LayerNorm, and pre-activation ResNets with Euclidean normalization. Each architecture has nine runs (three dimensions, three seeds). Each run retains the checkpoint with the lowest validation loss. Across 108 native test conditions, the mean absolute differences between predicted and observed median MSE reductions are 0.0238, 0.0126, and 0.00345, respectively. For the nine widest models at $q = 0.65$, eight independent copies multiply the relative RMSE to the analytic target by 0.3512 to 0.3532, close to $8^{-1/2} = 0.3536$, while shared masks preserve it. This illustrates Theorem 7, without establishing an asymptotic rate. Protocols and controls appear in Appendices E.2, F.3, F.7, and D.2. Table 2 reports the separate eight-block study.

Ten pretrained language models. On WikiText-2, we test ten pretrained GPT-2, OPT, T5, and BART checkpoints (Merity et al., 2017; Radford et al., 2019; Zhang et al., 2022; Raffel et al., 2020; Lewis et al., 2020). We resample only the last feed-forward residual branch before final normalization and vocabulary projection, fixing earlier activations (Section 3). Across five keep probabilities, predicted and observed condition medians of relative-MSE reductions have correlation 0.9834 and MAE 6.31×10^{-4}. At the native keep probability, the model at fixed radius predicts relative MSE reductions of $0.015\,\%$ to $0.362\,\%$, so residual dispersion dominates the scalar shift in this model. Pooling errors over 1 200 test tokens, the analytic prediction has $26.5\,\%$ lower MSE than the constant baselines given by each model's mean observed reduction on validation tokens (Appendix F.5).

For this local intervention, $\mathrm{softmax}(\ell_T)$ has $2.6\,\%$ to $66.5\,\%$ lower L1 error than $\mathrm{softmax}(\ell_0)$ against an independent MC-2048 probability target. In every model, bootstrapping passages gives a positive 95 % interval for the absolute gain, conditional on this fixed finite target. Averaging probabilities after softmax reduces Monte Carlo variability, not the discrepancy between their mean and $\mathrm{softmax}(\ell_T)$. In a separate test activating all native dropout operations, the fitted scalar reduces the complete centered-logit MSE by at most $16.7\,\%$ in median in each model, relative to the output without dropout. Protocols and additional results appear in Appendices E.3, F.4, and F.7.

Pretrained FaceNet embedding. FaceNet tests the approximation with 1 792 independent dropout sources feeding a 512-dimensional normalized embedding, with cosine-score verification on all 6 000 official LFW pairs (Schroff et al., 2015; Cao et al., 2018; Huang et al., 2007).

The MSE test instead compares 4 564 query images with a fixed 512-image gallery, giving a pooled relative one-pass MSE reduction of $1.34\,\%$ at the native keep probability. Separately, on 512 balanced official pairs, the analytic score reduces MAE to an independent MC-2048 target by $91.9\,\%$ relative to the deterministic score and outperforms all three tested sampled baselines in MAE. Thus single-pass dispersion can dominate despite accurate approximation of the Monte Carlo mean. For the rectangular theory, protocols, and MC-2048 results, see Appendices E.4, F.6, and F.6.1.

8 Conclusion

Under the studied conditions, normalization can make dropout approximately scalar when the radius concentrates and little numerator variation reaches the output. The identity at fixed radius, T, $\mathcal{V}$, and Λ separate scalar shift from residual dispersion. Radius concentration alone is insufficient. Experiments distinguish accurate approximation of the mean from residual variability and largely non-scalar changes throughout the network.

AI USE STATEMENT

Large language models were used to improve the clarity, grammar, and readability of the manuscript and to assist with debugging, and improving the clarity of experimental code. They were not used to formulate the research questions or hypotheses, develop the conceptual or theoretical contributions, derive mathematical results, or interpret experimental results.

REPRODUCIBILITY STATEMENT

All main assumptions and central theorem statements appear in the main paper. Appendix B provides the complete proofs. Appendix E gives the architectures details, data splits, mask counts, metrics, and aggregation rules. The supplementary `code_dropout.zip` contains the experiment scripts. Its `README.md` gives installation instructions, commands to regenerate the measurements.

REFERENCES

Raman Arora, Peter Bartlett, Poorya Mianjy, and Nathan Srebro. Dropout: Explicit forms and capacity control. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 351–361. PMLR, 2021.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Pierre Baldi and Peter J. Sadowski. Understanding dropout. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2814–2822. Curran Associates, Inc., 2013.

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *13th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 67–74, 2018.

Tianqi Chen, Ian Goodfellow, and Jonathon Shlens. Net2Net: Accelerating learning via knowledge transfer. In *International Conference on Learning Representations*, 2016.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059. PMLR, 2016.

Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3581–3590. Curran Associates, Inc., 2017.

Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V. Le. DropBlock: A regularization method for convolutional networks. In *Advances in Neural Information Processing Systems*, volume 31, pp. 10727–10737. Curran Associates, Inc., 2018.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. In *Computer Vision – ECCV 2016*, pp. 646–661. Springer, 2016. doi: 10.1007/978-3-319-46493-0_39.

Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

Bum Jun Kim, Hyeyeon Choi, Hyeonah Jang, Donggeon Lee, and Sang Woo Kim. How to use dropout correctly on residual networks with batch normalization. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 1058–1067. PMLR, 2023.

Diederik P. Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2575–2583. Curran Associates, Inc., 2015.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Benjamin J. Lengerich, Eric Xing, and Rich Caruana. Dropout as a regularizer of interaction effects. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 7550–7564. PMLR, 2022.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.703.

Xiang Li, Shuo Chen, Xiaolin Hu, and Jian Yang. Understanding the disharmony between dropout and batch normalization by variance shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2682–2690, 2019.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. R-Drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 10890–10905. Curran Associates, Inc., 2021.

Elizabeth S. Meckes. *The Random Matrix Theory of the Classical Compact Groups*. Cambridge University Press, 2019. doi: 10.1017/9781108303453.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *International Conference on Learning Representations*, 2017.

Poorya Mianjy and Raman Arora. On dropout and nuclear norm regularization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4575–4584. PMLR, 2019.

Poorya Mianjy and Raman Arora. On convergence and generalization of dropout training. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21151–21161. Curran Associates, Inc., 2020.

Poorya Mianjy, Raman Arora, and René Vidal. On the implicit bias of dropout. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3540–3548. PMLR, 2018.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Mark Rudelson and Roman Vershynin. Hanson–wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18(82):1–9, 2013. doi: 10.1214/ECP.v18-2865.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Stefan Wager, Sida Wang, and Percy S. Liang. Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using DropConnect. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1058–1066. PMLR, 2013.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. CosFace: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5265–5274, 2018.

Sida Wang and Christopher Manning. Fast dropout training. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 118–126. PMLR, 2013.

Colin Wei, Sham Kakade, and Tengyu Ma. The implicit and explicit regularization effects of dropout. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10181–10192. PMLR, 2020.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

APPENDIX CONTENTS

A Notation guide **14**

Proof roadmap **17**

B Complete proofs **19**

B.1 Concentration tools . . . 19
B.2 Proofs for the scalar construction and Section 4 . . . 19
B.3 Proofs for the law relating the mean shift to dispersion . . . 22
B.4 An explicit bound for the error in the predicted MSE reduction . . . 24
B.5 Proofs for replication preserving the original function . . . 25
B.6 Proof of the limiting distribution . . . 29
B.7 Proofs for basis dependence . . . 31

C Further implementation remarks **33**

D Several nonlinear layers **34**

D.1 Extension to deeper ReLU layers . . . 34
D.2 Experiments with several ReLU layers . . . 39

E Experimental details **42**

E.1 Common estimators . . . 42
E.2 MNIST architecture and training details . . . 43
E.3 Pretrained language model protocol . . . 44
E.4 Pretrained angular face-embedding protocol . . . 45

F Additional results **46**

F.1 Predicted and observed MSE reductions . . . 47
F.2 Empirical scalar comparisons . . . 47
F.3 Additional experiments on MNIST models . . . 49
F.4 Additional experiments on ten language models . . . 51
F.5 Predicting relative MSE reductions on held-out tokens . . . 56
F.6 Additional experiments on FaceNet . . . 57
F.6.1 Independent Monte Carlo reference for FaceNet . . . 59
F.7 Further MNIST and language model boundary tests . . . 60

A NOTATION GUIDE

The argument is local to one input, so the input and all network weights stay fixed while only the dropout mask is random. The tables below gather the symbols used in the statements and proofs.

Dimensions, representations, and maps

Symbol	Meaning
p, d, m	Feature dimension after normalization, number of independent random variables, and number of outputs. For coordinate dropout, $d = p$.
$x, u, h = x + u$	Part left unchanged, part to which dropout is applied, and representation without dropout in $\mathbb{R}^p$.
q, ρ	Probability of keeping a coordinate and positive variance of the centered mask. Nontrivial inverted Bernoulli dropout has $q \in \,]0, 1[$ and $\rho = (1-q)/q$. The case $q = 1$ is deterministic and has no dropout.
ζ, ξ	Centered mask and standardized mask, with $\xi = \zeta/\sqrt{\rho}$. Their coordinates are independent, have mean zero, and satisfy $\|\xi_i\|_{\psi_2} \le K$.
K	Common upper bound on the sub-Gaussian norms of the standardized mask coordinates.
D, M	Matrix $D = \mathrm{diag}(u) \in \mathbb{R}^{p\times p}$ for coordinate dropout and matrix $M \in \mathbb{R}^{p\times d}$ for the general random change $h_\zeta = h + M\zeta$.
P	Matrix selecting the part that is normalized. It is I_p for RMS or Euclidean normalization and $I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p^\top$ for LayerNorm.
C_m	Output-centering matrix $I_m - m^{-1}\mathbf{1}_m\mathbf{1}_m^\top$ for softmax logits. For cosine scores, $C_m = I_m$.
A, c	Effective output matrix $A = \kappa C_m W \Gamma P \in \mathbb{R}^{m\times p}$ and fixed output vector $c = C_m(W\beta_{\mathrm{N}} + b) \in \mathbb{R}^m$, with $AP = A$, $C_m A = A$, and $C_m c = c$. The matrix W is the last linear weight, Γ and β_{N} are the normalization gain and shift, and κ converts the implementation normalization to the Euclidean radius convention.
$v, \mathcal{N}_{P,\varepsilon}, r(v)$	A generic representation $v \in \mathbb{R}^p$, its normalization map, and its radius, with $\mathcal{N}_{P,\varepsilon}(v) = Pv/r(v)$ and $r(v) = (\|Pv\|_2^2 + \varepsilon)^{1/2}$. For RMSNorm and LayerNorm, the fixed factor $\sqrt{p}$ is absorbed into A and ε is the rescaled implementation stabilizer.
$\ell(v), \ell_0, \ell_\zeta$	Output map $\ell(v) = c + A\mathcal{N}_{P,\varepsilon}(v)$, output without dropout $\ell_0 = \ell(h)$, and output for a drawn mask $\ell_\zeta = \ell(h + D\zeta)$.

Radius, scalar factors, and probability bounds	
Symbol	**Meaning**
$r_0, r_\zeta, \bar{r}$	Radius without dropout, radius for the drawn mask, and root-mean-square radius, which does not depend on the drawn mask. They satisfy $r_0 = r(h)$, $r_\zeta = r(h + D\zeta)$, and $\bar{r}^2 = \mathbb{E} r_\zeta^2 = r_0^2 + \rho \operatorname{Tr}(DPD)$.
T, ℓ_T	Radius ratio $T = \bar{r}/r_0$ and fixed target $\ell_T = c + (\ell_0 - c)/T = c + APh/\bar{r}$.
B	Matrix $B = DPD$ in the quadratic radius term. Its Frobenius and operator norms measure the aggregate and largest-direction contributions to radius fluctuations.
$\mathcal{V}$	Dropout variance reaching the output divided by the squared target that depends on the features, $\mathcal{V} = \rho \lVert APD \rVert_{\mathrm{F}}^2 / \lVert APh \rVert_2^2$, defined when $APh \neq 0$. It equals $D_{\mathrm{vis}} / \lVert \ell_T - c \rVert_2^2$, the residual MSE normalized by the squared target size, in the model with a fixed radius.
$\mathcal{V}_{\Sigma,k}$	Covariance-aware visible variance ratio for a replicated mask vector with covariance Σ_k, defined in (53). It equals $\mathcal{V}_k$ for independent masks and remains equal to $\mathcal{V}_1$ for shared masks.
r_{vis}	Stable rank $\lVert APD \rVert_{\mathrm{F}}^2 / \lVert APD \rVert_{\mathrm{op}}^2$. It measures the effective number of directions that send dropout noise to the output and is set to $+\infty$ when $APD = 0$.
t, θ	The parameter $t \in [1, +\infty[$ controls failure probability. The auxiliary tolerance $\theta \in]0, 1[$ is used only in the explicit mean-square bound of Appendix B.
$\mathfrak{R}_t, \mathfrak{D}_t$	Bounds for $\lvert r_\zeta^2 - \bar{r}^2 \rvert$ and $\lVert APD\zeta \rVert_2$.
χ_t	Relative tolerance for the squared radius $\chi_t = \mathfrak{R}_t / \bar{r}^2$.

Mean shift and dispersion at fixed radius	
Symbol	**Meaning**
$\widetilde{\ell}_\zeta$	Output obtained by replacing the random radius with $\bar{r}$, namely $\widetilde{\ell}_\zeta = c + (APh + APD\zeta)/\bar{r}$.
$\mathcal{Q}_{\mathrm{fr}}(a), a_\star$	Mean squared error of the scalar model $c + a(\ell_0 - c)$ when the radius is fixed, and its unique best scalar $a_\star = T^{-1}$ when $APh \neq 0$.
$R_{\mathrm{shift}}, D_{\mathrm{vis}}$	Squared systematic shift $\lVert \ell_T - \ell_0 \rVert_2^2$ and stochastic dispersion $\mathbb{E} \lVert \widetilde{\ell}_\zeta - \ell_T \rVert_2^2$. Their explicit values are $(T-1)^2 \lVert APh \rVert_2^2 / \bar{r}^2$ and $\rho \lVert APD \rVert_{\mathrm{F}}^2 / \bar{r}^2$.
$\mathcal{E}(L)$	Relative MSE reduction obtained by using ℓ_T instead of ℓ_0 for the same square-integrable random output L. It is defined when $\mathbb{E} \lVert L - \ell_0 \rVert_2^2 > 0$.
Λ	Ratio of the removable mean shift to the residual dispersion, $\Lambda = R_{\mathrm{shift}} / D_{\mathrm{vis}} = (T-1)^2 / \mathcal{V}$, defined when $D_{\mathrm{vis}} > 0$. In the model with a fixed radius, $\mathcal{E} = \Lambda / (1 + \Lambda)$.
$\ell_\zeta, \mathcal{Q}(a), a_{\mathrm{true}}$	Original normalized output, its mean squared scalar error, and its best scalar.
$\epsilon_{\mathrm{fr}}, \eta_{\mathrm{fr}}$	Root mean squared difference between ℓ_ζ and the output at fixed radius, and this error relative to the complete change at fixed radius. They satisfy $\epsilon_{\mathrm{fr}}^2 = \mathbb{E} \lVert \ell_\zeta - \widetilde{\ell}_\zeta \rVert_2^2$ and, when the denominator is positive, $\eta_{\mathrm{fr}} = \epsilon_{\mathrm{fr}} / \sqrt{R_{\mathrm{shift}} + D_{\mathrm{vis}}}$.

Rectangular and deep extensions	
Symbol	**Meaning**
$B_M, \bar{r}_M, T_M$	Rectangular-model radius matrix $B_M = M^\top PM$, reference radius $\bar{r}_M^2 = r_0^2 + \rho \operatorname{Tr}(B_M)$, and radius ratio $T_M = \bar{r}_M / r_0$.
$\mathcal{V}_M, \Lambda_M, \mathcal{E}_M$	Rectangular analogues of the relative residual MSE, shift-to-dispersion ratio, and relative MSE reduction. The ratio $\mathcal{V}_M$ requires $APh \neq 0$, $\Lambda_M = (T_M - 1)^2/\mathcal{V}_M$ requires $\mathcal{V}_M > 0$, and the target is $\ell_{T_M} = c + (\ell_0 - c)/T_M$.
$J_\ell, \mathcal{M}_t, M_{\text{eff}}$	Selected local linear map of ReLU block ℓ, matrix propagating all random sources to layer t, and its final value $M_{\text{eff}} = \mathcal{M}_L$. When all preactivations are nonzero, J_ℓ is the usual Jacobian.
$g_\zeta, g_\zeta^{\text{tot}}, \delta_{\text{gate}}$	Exact nonlinear term created by ReLU state changes in the additive model, total residual including mask interactions in the layerwise experiment, and the upper bound on the probability that at least one state changes.
$\epsilon_{\text{gate}}, \epsilon_{\text{fr},M}$	Root mean squared logit errors due to ReLU state changes and to replacing the random radius by $\bar{r}_M$. Their sum is $\epsilon_{\text{deep},M}$.
$\eta_M, \widehat{G}$	Relative deep-approximation error used in the stability theorem and empirical ratio of the total residual to the linear perturbation used in the deep experiment.

Replication and changes of coordinates	
Symbol	**Meaning**
k, L_k	Number of copies and map $L_k \colon \mathbb{R}^p \longrightarrow \mathbb{R}^{kp}$, which replaces each coordinate with k copies divided by $\sqrt{k}$.
h_k, u_k, A_k	Replicated quantities $h_k = L_k h$, $u_k = L_k u$, and $A_k = AL_k^\top$.
$\mathcal{V}_k$	Numerator-noise-to-signal ratio $\mathcal{V}$ from (15), recomputed in the k-copy model using (h_k, u_k, A_k, P_{kp}). Thus $\mathcal{V}_1$ is the value before replication.
d_{eff}, μ	Effective number of coordinate contributions that reach the output and their largest relative share, defined when the perturbation reaching the output is nonzero.
Q, q_i	Orthogonal change of coordinates $Q \in O(p)$ and its vectors $q_i = Q^\top e_i$, where e_i is the ith standard basis vector. Scalar risks use the distinct symbol $\mathcal{Q}$.
Conventions	
$\|\cdot\|_2, \|\cdot\|_{\text{F}}, \|\cdot\|_{\text{op}}$	Euclidean, Frobenius, and operator norms.
$\mathbb{E}, \mathbb{P}, \operatorname{Tr}, \operatorname{rank}$	Expectation, probability, trace, and matrix rank. Unless stated otherwise, expectation and probability refer to the dropout mask.
$O_{\mathbb{P}}(\cdot), \xrightarrow{d}$	Order in probability and convergence in distribution. The symbol $C > 0$ denotes the fixed absolute constant used in the named concentration bounds. Unnamed constants may change between formulas.

PROOF ROADMAP

The two tables below connect the results to their proof ideas. The preceding notation guide defines the symbols. Throughout the mask arguments, the input, branch values, and trained weights are fixed. The target $\ell_T = c + (\ell_0 - c)/T$ rescales only the part that depends on the features while keeping c fixed. It is not generally ℓ_0/T. An identity for the model with radius $\bar{r}$ is distinguished from a bound for the actual model with a random radius.

CORE ARGUMENT: FROM RANDOM NORMALIZATION TO SCALAR APPROXIMATION

Result	Key proof idea	Assumptions
Radius and numerator split Proposition 1 *Exact for every mask.*	Separate the numerator into APh and $APD\zeta$, and factor r_0/r_ζ out of the first term. Subtracting ℓ_T gives a scalar-coefficient error plus the normalized numerator perturbation $APD\zeta/r_\zeta$.	This is an algebraic identity in the stated normalized model. The numerator term may have a component along $\ell_0 - c$: the split is not orthogonal for each mask.
Finite-dimensional control Theorem 2 and Corollary 3 *High-probability bound and sufficient asymptotic condition.*	Expand $r_\zeta^2 - \bar{r}^2$ into a linear term and a centered quadratic form. Lemma 10 bounds these terms and the visible numerator. A lower bound on r_ζ then controls its inverse. Combine the radius and numerator errors by the triangle inequality.	Independent sub-Gaussian masks and $\chi_t < 1$ give the normalized bound. Relative error requires $APh \neq 0$. Radius concentration alone is insufficient: the visible-noise term must also be small. The corollary states sufficiency, not necessity.
Best scalar and relative MSE reduction Theorem 4 and Corollary 5 *Exact in mean square for the model with a fixed radius.*	The numerator term at fixed radius has mean zero, so the expected cross term vanishes, giving $\mathcal{Q}_{\mathrm{fr}}(a) = (a - T^{-1})^2 \|\ell_0 - c\|_2^2 + D_{\mathrm{vis}}$. This identity relating the mean shift to dispersion gives $a_\star = T^{-1}$. Comparing the baseline risk at $a = 1$ with the minimum at $a = 1/T$ gives $\mathcal{E} = \Lambda/(1 + \Lambda)$.	$APh \neq 0$ ensures uniqueness. The corollary assumes $\mathcal{V} > 0$ to define Λ. The reduction itself only needs a positive baseline MSE. Orthogonality is in L^2 over masks. These are exact identities for $\widetilde{\ell}_\zeta$, not automatically for the actual output ℓ_ζ.
Effect of restoring the random radius Proposition 6 *Perturbation bounds in mean square.*	Project $\mathbb{E}[\ell_\zeta - c]$ onto $\ell_0 - c$ to obtain the best scalar coefficient. Cauchy–Schwarz bounds its displacement from T^{-1} by $\epsilon_{\mathrm{fr}}/\|\ell_0 - c\|_2$. Bound the changes in both mean squared errors before comparing their ratio in $\mathcal{E}$.	The scalar bounds require $\ell_0 - c \neq 0$ and the stated L^2 error control. The relative MSE reduction bound additionally needs $\mathcal{V} > 0$ and $\eta_{\mathrm{fr}} \leq 1/4$. A small absolute error alone need not control a ratio whose reference denominator is small.
An explicit L^2 error bound Corollary 12 *Finite-dimensional bound.*	On the concentration event, the error from fixing the radius is at most g_t. On its complement, bounded masks give a uniform bound B_b. Split the expectation to obtain $\epsilon_{\mathrm{fr}}^2 \leq g_t^2 + 6e^{-t}B_b^2$, then apply the preceding perturbation result.	Requires bounded masks and $\mathfrak{R}_t \leq \theta\bar{r}^2$. Inverted Bernoulli dropout has $\|\zeta_i\| \leq \max\{1, \rho\}$. The exceptional-event term is needed: a high-probability estimate alone does not establish the claimed mean-square control.

CONSTRUCTIONS AND EXTENSIONS

Result and status	Key proof idea	Assumptions
Independent noise through a linear map Proposition 11 *Same identities and bounds for $h + M\zeta$.*	Replace the linear radius vector by $M^\top Ph$, the quadratic matrix by $M^\top PM$, and the visible-noise matrix by APM. The concentration, inverse-radius, and second-moment arguments then apply with these substitutions.	M is deterministic. The source coordinates of ζ are independent, centered, and sub-Gaussian as stated. Coordinates of $M\zeta$ need not be independent. This result does not directly allow M to depend on the mask.
Replication and shared masks Propositions 13 and 14 *Exact function and maskwise identities.*	Use $L_k^\top L_k = I_p$ and $P_{kp}L_k = L_kP_p$ to commute replication with normalization. Under shared masks the perturbed representation is exactly $L_k(h + D\zeta)$, so mapping back preserves even the random output.	The specified $1/\sqrt{k}$ replication and matching output map are essential. Both RMS and LayerNorm are covered. More stored coordinates alone give no reduction when every copy shares its original mask.
Independent-copy rate Theorem 7 *Exact variance scaling and asymptotic error bound.*	Each original visible-noise column is replaced by k columns divided by k. Thus the visible-numerator covariance and $\mathcal{V}$ are divided by k. The radius bound scales as $k^{-1/2}$. Inserting both scalings into the main result gives the same output-error upper bound.	Fix the base problem. The copy masks are independent and centered, have common variance ρ, and satisfy a sub-Gaussian bound uniform in k. The visible stable rank stays unchanged. LayerNorm adds an $O(k^{-1})$ centering correction.
Gaussian limit and MC approximation Theorem 15 and Corollary 16 *Asymptotic distribution and expectation bound.*	Apply a joint central limit theorem to averages of masks and squared masks, then the delta method to normalization. For Bernoulli masks, a second-order expansion of softmax after normalization has a centered linear term. A bounded Hessian gives a mean-prediction error of order k^{-1}.	The Gaussian limit is for RMS replication with a finite fourth moment. A nonzero Σ_∞ rules out a faster rate. The MC bound uses Bernoulli masks and $\varepsilon > 0$. Both concern the replicated model, not a generic fixed pretrained network.
Basis dependence Theorem 8, Propositions 9 and 17 *Exact invariance, Haar average, and basis construction.*	Orthogonal reparameterization preserves the baseline and T. Spherical fourth moments give the factor $1/(p+2)$ in the Haar average of visibility. For a contrasting basis, span the row space of A first, then complete with its kernel. Only $r_A = \mathrm{rank}(A)$ axes can contribute directly to the visible numerator.	The Haar formula and aligned-basis construction assume $P = I_p$. The LayerNorm result is stated for the sufficient subgroup $Q\mathbf{1}_p = \mathbf{1}_p$. The expectation is over Q, not masks. It is not an every-basis guarantee. Concentrating the support alone gives no lower bound on $\mathcal{V}$.
ReLU propagation and state changes Theorem 18 and Proposition 19 *Exact split and probability bound.*	Unroll the layer recursion into propagated linear changes $M_{\text{eff}}\zeta$ plus ReLU remainders g_ζ. Before the first state change, propagation is exactly linear. A sub-Gaussian tail bound for each possible first change and a union bound control its probability.	The model uses fixed additive noise matrices and independent source coordinates. The probability bound assumes positive ReLU margins and can be uninformative if many units can switch. With unchanged states, $g_\zeta = 0$ exactly.
Scalar approximation after ReLU layers Theorem 20 *Mean-square perturbation bounds.*	Add the L^2 error from ReLU changes to the error from fixing the radius. The rectangular model with a fixed radius supplies the reference optimum and relative MSE reduction. Proposition 6 bounds their change when the random radius and ReLU remainders are restored.	Requires $APh \neq 0$ and $\mathcal{V}_M > 0$. The relative MSE reduction bound also needs $\eta_M \leq 1/4$. Standard multilayer dropout introduces additional terms that depend on the mask, which must enter the error control. The explicit gate probability concerns the fixed additive-noise model only.

B COMPLETE PROOFS

The proofs follow the order of the main text. We first combine linear, Hanson-Wright, and matrix concentration to bound the radius and the numerator. An orthogonal squared-norm identity then yields the law relating the mean shift to dispersion. Exact matrix scaling explains the replication experiment, a multivariate central limit theorem gives a condition that rules out a faster rate, and a spherical fourth-moment calculation gives the average over random coordinate systems.

B.1 CONCENTRATION TOOLS

We first record three standard consequences of sub-Gaussian concentration. The fixed absolute constant C introduced in Section 4 is chosen large enough that all three bounds below hold.

Lemma 10 (Linear, quadratic, and matrix deviations)**.** *Let $z = (z_1, \dots, z_p)$ have independent, centered entries with $\mathbb{E}z_i^2 = 1$ and $\|z_i\|_{\psi_2} \le K$. For every $t \in [1, +\infty[$, the following events hold with the stated probabilities:*

$$\left|a^\top z\right| \le CK \|a\|_2 \sqrt{t}, \qquad \text{with probability at least } 1 - 2e^{-t}, \tag{39}$$

$$\left|z^\top B z - \operatorname{Tr}(B)\right| \le CK^2 \left(\|B\|_{\mathrm{F}} \sqrt{t} + \|B\|_{\mathrm{op}} t\right), \qquad \text{with probability at least } 1 - 2e^{-t}, \tag{40}$$

$$\|Mz\|_2 \le CK^2 \left(\|M\|_{\mathrm{F}} + \|M\|_{\mathrm{op}} \sqrt{t}\right), \qquad \text{with probability at least } 1 - 2e^{-t}. \tag{41}$$

The vector a is deterministic, B is deterministic and symmetric, and M is a deterministic matrix of any output dimension.

Equation (40) follows from the Hanson–Wright inequality (Rudelson & Vershynin, 2013, Theorem 1.1). For (41), Theorem 2.1 of the same reference controls $\|Mz\|_2$ around $\|M\|_{\mathrm{F}} = (\mathbb{E}\|Mz\|_2^2)^{1/2}$. See also Vershynin (2018). We use the slightly loose factor K^2 in (41) so that one constant covers all $K \ge 1$.

B.2 PROOFS FOR THE SCALAR CONSTRUCTION AND SECTION 4

Proof of Proposition 1. By (1) and (4), the random output is

$$\begin{aligned}
\ell_\zeta &= c + \frac{AP(h + D\zeta)}{r_\zeta} \\
&= c + \frac{r_0}{r_\zeta}\frac{APh}{r_0} + \frac{APD\zeta}{r_\zeta} \\
&= c + \frac{r_0}{r_\zeta}(\ell_0 - c) + \frac{APD\zeta}{r_\zeta}.
\end{aligned}$$

The last equality uses $\ell_0 - c = APh/r_0$. This proves (5). The analytic target is $\ell_T = c + (\ell_0 - c)/T$, so subtracting it gives

$$\ell_\zeta - \ell_T = \left(\frac{r_0}{r_\zeta} - \frac{1}{T}\right)(\ell_0 - c) + \frac{APD\zeta}{r_\zeta},$$

which also gives (9). □

Proof of Theorem 2. If $\rho = 0$, then $\zeta = 0$ almost surely and every stated bound is immediate. Suppose $\rho > 0$. Set $z = \zeta/\sqrt{\rho}$. Its coordinates are independent, centered, have unit variance, and have sub-Gaussian norm at most K. Since P is a symmetric projector and D is diagonal, expanding the random squared radius gives

$$\begin{aligned}
r_\zeta^2 &= \|P(h + D\zeta)\|_2^2 + \varepsilon \\
&= r_0^2 + 2(DPh)^\top \zeta + \zeta^\top DPD\zeta.
\end{aligned} \tag{42}$$

Because $\mathbb{E}\zeta\zeta^\top = \rho I_p$,

$$\mathbb{E}\left[\zeta^\top DPD\zeta\right] = \rho \operatorname{Tr}(DPD).$$

Thus $\bar{r}^2 = r_0^2 + \rho \operatorname{Tr}(DPD)$ is exactly the mean of r_ζ^2. Lemma 10 requires independent, centered, unit-variance sub-Gaussian coordinates and deterministic coefficient matrices, which are precisely the properties of z and the matrices below. Its linear estimate controls the scalar cross term, while Hanson–Wright controls the centered quadratic term. Apply (39) to the linear form with $a = DPh$ and (40) to the quadratic form with $B = DPD$. The union bound has the form $\mathbb{P}(\bigcup_j E_j) \le \sum_j \mathbb{P}(E_j)$ and requires no independence between the events. Applying it to the two failure events, restoring the factors $\sqrt{\rho}$ and ρ, and absorbing the factor two into the absolute constant yields

$$\left| r_\zeta^2 - \bar{r}^2 \right| \le \mathfrak{R}_t \tag{43}$$

with probability at least $1 - 4e^{-t}$. Separately, applying (41) to $M = APD$ gives $\|APD\zeta\|_2 \le \mathfrak{D}_t$ with probability at least $1 - 2e^{-t}$. A final union bound therefore gives both events in (18) with probability at least $1 - 6e^{-t}$.

Suppose now that (43) holds and $\chi_t < 1$. Since $\mathfrak{R}_t = \chi_t \bar{r}^2$, the lower side of the bound on the squared radius gives

$$r_\zeta \ge \sqrt{1 - \chi_t}\, \bar{r}. \tag{44}$$

To compare inverse radii, use $|r_\zeta - \bar{r}| = |r_\zeta^2 - \bar{r}^2|/(r_\zeta + \bar{r})$. Hence

$$\begin{aligned} \left| \frac{r_0}{r_\zeta} - \frac{1}{T} \right| &= r_0 \left| \frac{1}{r_\zeta} - \frac{1}{\bar{r}} \right| \\ &= \frac{r_0}{\bar{r}} \frac{|r_\zeta^2 - \bar{r}^2|}{r_\zeta (r_\zeta + \bar{r})} \\ &\le T^{-1} \frac{\chi_t}{\sqrt{1-\chi_t}(1 + \sqrt{1-\chi_t})}, \end{aligned}$$

which is (19). Equations (44) and $\|APD\zeta\|_2 \le \mathfrak{D}_t$ also give

$$\left\| \frac{APD\zeta}{r_\zeta} \right\|_2 = \frac{\|APD\zeta\|_2}{r_\zeta} \le \frac{\mathfrak{D}_t}{\sqrt{1-\chi_t}\, \bar{r}},$$

which is (20).

It remains to express both estimates relative to the component that depends on the features and that the scalar approximation seeks to reproduce. Since $APh \neq 0$,

$$\left\| \frac{\ell_0 - c}{T} \right\|_2 = \frac{\|APh\|_2}{\bar{r}}.$$

The scalar part of (9), divided by this quantity, is

$$T \left| \frac{r_0}{r_\zeta} - \frac{1}{T} \right| \le \frac{\chi_t}{\sqrt{1-\chi_t}(1 + \sqrt{1-\chi_t})}.$$

If $APD = 0$, the visible numerator term vanishes. Otherwise, the definitions of $\mathcal{V}$ and r_{vis} give

$$\begin{aligned} \frac{\mathfrak{D}_t}{\sqrt{1-\chi_t}\, \|APh\|_2} &\le \frac{CK^2}{\sqrt{1-\chi_t}} \frac{\sqrt{\rho}\, \|APD\|_{\mathrm{F}}}{\|APh\|_2} \left(1 + \frac{\|APD\|_{\mathrm{op}}}{\|APD\|_{\mathrm{F}}} \sqrt{t} \right) \\ &= \frac{CK^2}{\sqrt{1-\chi_t}} \sqrt{\mathcal{V}} \left(1 + \sqrt{\frac{t}{r_{\mathrm{vis}}}} \right). \end{aligned}$$

The triangle inequality in (9) and the last two displays prove (21). □

Proof of Corollary 3. Set $t = \log(6/\delta)$ as in the statement, so $t > 1$ and $1 - 6e^{-t} = 1 - \delta$. Since $\chi_{t,n} \to 0$, we have $\chi_{t,n} < 1$ for all sufficiently large n. The first term in (21) then converges to zero.

For the second term, write $x_n = t/r_{\mathrm{vis},n}$. The elementary inequality

$$(1 + \sqrt{x_n})^2 \le 2(1 + x_n)$$

shows that

$$K_n^2 \sqrt{\mathcal{V}_n}(1 + \sqrt{x_n}) \le \left(\sup_m K_m^2 \right) \sqrt{2\mathcal{V}_n(1 + x_n)} \longrightarrow 0.$$

The uniform bound on K_n controls the constant multiplying the visible-numerator term. Together with (23), this shows that the complete right-hand side of (21) is eventually at most any prescribed $\epsilon > 0$, on an event of probability at least $1 - \delta$. □

For a generic representation $v \in \mathbb{R}^p$, write $r(v) = (\|Pv\|_2^2 + \varepsilon)^{1/2}$ and $\ell(v) = c + A\mathcal{N}_{P,\varepsilon}(v) = c + APv/r(v)$.

Proposition 11 (Rectangular linear-image extension)**.** *Assume $APh \neq 0$. Let $\zeta \in \mathbb{R}^d$ have independent, centered entries with variance $\rho > 0$ and $\|\zeta_j/\sqrt{\rho}\|_{\psi_2} \leq K$. For a deterministic $M \in \mathbb{R}^{p\times d}$, set $h_\zeta = h + M\zeta$ and*

$$B_M = M^\top PM, \qquad \bar{r}_M^2 = r_0^2 + \rho \operatorname{Tr}(B_M),$$
$$\mathcal{V}_M = \frac{\rho\|APM\|_{\mathrm{F}}^2}{\|APh\|_2^2}, \qquad r_{\mathrm{vis},M} = \frac{\|APM\|_{\mathrm{F}}^2}{\|APM\|_{\mathrm{op}}^2}.$$

As before, $r_{\mathrm{vis},M} = +\infty$ when $APM = 0$. Define

$$\mathfrak{R}_{M,t} = C\left[K\sqrt{\rho}\,\|M^\top Ph\|_2\sqrt{t} + K^2\rho\left(\|B_M\|_{\mathrm{F}}\sqrt{t} + \|B_M\|_{\mathrm{op}}t\right)\right],$$
$$\mathfrak{D}_{M,t} = CK^2\sqrt{\rho}\left(\|APM\|_{\mathrm{F}} + \|APM\|_{\mathrm{op}}\sqrt{t}\right).$$

Then Proposition 1 and Theorem 2 hold after replacing $D\zeta$ by $M\zeta$, DPh by $M^\top Ph$, DPD by B_M, APD by APM, $\bar{r}$ by $\bar{r}_M$, $\mathfrak{R}_t$ by $\mathfrak{R}_{M,t}$, and $\mathfrak{D}_t$ by $\mathfrak{D}_{M,t}$ in all dependent quantities. Write $T_M = \bar{r}_M/r_0$, $\ell_{T_M} = c + (\ell_0 - c)/T_M$, and $\widetilde{\ell}_{\zeta,M} = c + (APh + APM\zeta)/\bar{r}_M$. Its optimal scale at fixed radius is $a_{\star,M} = T_M^{-1}$, and

$$\frac{\min_a \mathbb{E}\|\widetilde{\ell}_{\zeta,M} - \{c + a(\ell_0 - c)\}\|_2^2}{\|a_{\star,M}(\ell_0 - c)\|_2^2} = \mathcal{V}_M.$$

Define the rectangular shift and dispersion by

$$R_{\mathrm{shift},M} = \frac{(T_M - 1)^2\|APh\|_2^2}{\bar{r}_M^2}, \qquad D_{\mathrm{vis},M} = \frac{\rho\|APM\|_{\mathrm{F}}^2}{\bar{r}_M^2}.$$

For any square-integrable random output L satisfying $\mathbb{E}\|L - \ell_0\|_2^2 > 0$, define

$$\mathcal{E}_M(L) = \frac{\mathbb{E}\|L - \ell_0\|_2^2 - \mathbb{E}\|L - \ell_{T_M}\|_2^2}{\mathbb{E}\|L - \ell_0\|_2^2}.$$

If $\mathcal{V}_M > 0$, the corresponding law relating the mean shift to dispersion is

$$\mathcal{E}_M(\widetilde{\ell}_{\zeta,M}) = \frac{\Lambda_M}{1 + \Lambda_M}, \qquad \Lambda_M = \frac{(T_M - 1)^2}{\mathcal{V}_M}.$$

If $\mathcal{V}_M = 0$, then $D_{\mathrm{vis},M} = 0$. If $R_{\mathrm{shift},M} > 0$, the scalar target removes the complete MSE at fixed radius and $\mathcal{E}_M(\widetilde{\ell}_{\zeta,M}) = 1$. If both terms vanish, the baseline MSE is zero and $\mathcal{E}_M$ is undefined. The ratio Λ_M is undefined in both cases. For the true normalized output $X_M = \ell(h + M\zeta)$, define

$$\epsilon_{\mathrm{fr},M}^2 = \mathbb{E}\|X_M - \widetilde{\ell}_{\zeta,M}\|_2^2,$$
$$\eta_{\mathrm{fr},M} = \frac{\epsilon_{\mathrm{fr},M}}{\sqrt{R_{\mathrm{shift},M} + D_{\mathrm{vis},M}}}.$$

The last quantity is defined when its denominator is positive. The stability bounds of Proposition 6 hold with these rectangular quantities. In particular, when $\mathcal{V}_M > 0$ and $\eta_{\mathrm{fr},M} \leq 1/4$, its relative MSE reduction bound applies to $\mathcal{E}_M(X_M)$.

Proof of Proposition 11. Set $\xi = \zeta/\sqrt{\rho}$. The coordinates of ξ satisfy the hypotheses of Lemma 10. Using the symmetry and idempotence of P, the squared radius has the exact expansion

$$r(h + M\zeta)^2 - \bar{r}_M^2 = 2\sqrt{\rho}\,\langle M^\top Ph, \xi\rangle + \rho\{\xi^\top B_M\xi - \operatorname{Tr}(B_M)\}.$$

The linear, quadratic, and matrix parts of Lemma 10 apply respectively to $M^\top Ph$, $B_M = M^\top PM$, and APM. The lemma requires deterministic coefficient matrices and independent centered standardized coordinates, so its hypotheses hold by the definition of ξ. The three estimates control the cross term, the centered squared radius, and the visible numerator in that order. They give $\mathfrak{R}_{M,t}$ and

$\mathfrak{D}_{M,t}$ exactly as stated. Once the squared radius is bounded below, the reciprocal identity used in the proof of Theorem 2 gives the same bounds for the normalized output with $D\zeta$ replaced by $M\zeta$.

For the conclusions at fixed radius, consider the random numerator term $APM\zeta/\bar{r}_M$. For any centered random vector w with finite covariance Σ and any deterministic matrix C, the covariance identity $\mathbb{E}\,\|Cw\|_2^2 = \mathrm{Tr}(C\Sigma C^\top)$ converts a second moment into a matrix trace. Since ζ is centered with covariance ρI_d, we obtain

$$\mathbb{E}\left[\frac{APM\zeta}{\bar{r}_M}\right] = 0, \qquad \mathbb{E}\left\|\frac{APM\zeta}{\bar{r}_M}\right\|_2^2 = \frac{\rho\,\|APM\|_{\mathrm{F}}^2}{\bar{r}_M^2}.$$

Therefore the deterministic displacement $(T_M^{-1} - a)(\ell_0 - c)$ is orthogonal in expectation to this random term. The projection calculation of Theorem 4 yields the unique optimum $a_{\star,M} = T_M^{-1}$ and the stated relative risk $\mathcal{V}_M$. When $\mathcal{V}_M > 0$, the same Pythagorean identity gives the law relating the mean shift to dispersion with the squared shift and residual dispersion defined above.

Finally, the proof of Proposition 6 depends only on the L^2 distance between the true output and its surrogate at fixed radius, the scalar line $c + a(\ell_0 - c)$, and those two reference terms. Applying its perturbation argument to X_M and $\widetilde{\ell}_{\zeta,M}$ gives the rectangular stability bounds. The condition $\eta_{\mathrm{fr},M} \le 1/4$ ensures that the denominator defining the actual relative MSE reduction remains positive, exactly as in that proof. □

B.3 Proofs for the law relating the mean shift to dispersion

The numerator perturbation $APD\zeta$ has mean zero. Dividing it by the fixed radius $\bar{r}$ preserves this property, so its expected inner product with the deterministic shift vanishes. This gives the mean squared error decomposition below. We then bound the effect of restoring the random radius.

Proof of Theorem 4. Recall from (24) that $\widetilde{\ell}_\zeta = \ell_T + APD\zeta/\bar{r}$, where $\ell_T = c + (\ell_0 - c)/T$. Thus, for any $a \in \mathbb{R}$,

$$\widetilde{\ell}_\zeta - \{c + a(\ell_0 - c)\} = (T^{-1} - a)(\ell_0 - c) + \frac{APD\zeta}{\bar{r}}.$$

The mask coordinates are centered, so $\mathbb{E}[APD\zeta/\bar{r}] = APD\mathbb{E}\zeta/\bar{r} = 0$. Consequently,

$$\mathbb{E}\left\langle (T^{-1} - a)(\ell_0 - c), \frac{APD\zeta}{\bar{r}} \right\rangle = 0.$$

The cross term vanishes because a deterministic vector paired with a centered random vector has zero expected inner product. More generally, for a centered vector ξ with finite covariance Σ, $\mathbb{E}\,\|C\xi\|_2^2 = \mathrm{Tr}(C\Sigma C^\top)$. Applying this covariance identity with $\mathbb{E}[\zeta\zeta^\top] = \rho I_p$ gives

$$\mathbb{E}\left\|\frac{APD\zeta}{\bar{r}}\right\|_2^2 = \frac{1}{\bar{r}^2}\,\mathrm{Tr}\big(APD\mathbb{E}[\zeta\zeta^\top]DPA^\top\big) = \frac{\rho\,\|APD\|_{\mathrm{F}}^2}{\bar{r}^2} = D_{\mathrm{vis}}.$$

Expanding the squared norm and using the zero cross term proves (26). Since $APh \neq 0$, we have $\ell_0 - c = APh/r_0 \neq 0$. The only a-dependent term is therefore a strictly convex quadratic, whose unique minimizer is $a_\star = T^{-1}$. At this value, $\mathcal{Q}_{\mathrm{fr}}(a_\star) = D_{\mathrm{vis}}$, which proves (28). Moreover,

$$\frac{\mathcal{Q}_{\mathrm{fr}}(a_\star)}{\|\ell_T - c\|_2^2} = \frac{\rho\,\|APD\|_{\mathrm{F}}^2/\bar{r}^2}{\|APh\|_2^2/\bar{r}^2} = \mathcal{V},$$

which proves (29).

Finally, setting $a = 1$ in (26) and using $(T^{-1} - 1)(\ell_0 - c) = \ell_T - \ell_0$ gives $\mathcal{Q}_{\mathrm{fr}}(1) = R_{\mathrm{shift}} + D_{\mathrm{vis}}$, which proves (27). □

Proof of Corollary 5. By definition of the surrogate at fixed radius,

$$\widetilde{\ell}_\zeta - \ell_T = \frac{APD\zeta}{\bar{r}}.$$

Theorem 4 therefore gives

$$\mathbb{E}\left\|\widetilde{\ell}_\zeta - \ell_T\right\|_2^2 = D_{\text{vis}},$$
$$\mathbb{E}\left\|\widetilde{\ell}_\zeta - \ell_0\right\|_2^2 = R_{\text{shift}} + D_{\text{vis}}.$$

Because $\mathcal{V} > 0$, we have $D_{\text{vis}} > 0$, so both the ratio below and the relative MSE reduction are well defined. From (25) and (15),

$$\frac{R_{\text{shift}}}{D_{\text{vis}}} = \frac{(T-1)^2 \|APh\|_2^2}{\rho \|APD\|_{\text{F}}^2} = \frac{(T-1)^2}{\mathcal{V}} = \Lambda.$$

Substituting R_{shift} and D_{vis} into (30) gives

$$\begin{aligned}\mathcal{E}(\widetilde{\ell}_\zeta) &= 1 - \frac{D_{\text{vis}}}{R_{\text{shift}} + D_{\text{vis}}} \\ &= \frac{R_{\text{shift}}/D_{\text{vis}}}{1 + R_{\text{shift}}/D_{\text{vis}}} = \frac{\Lambda}{1+\Lambda},\end{aligned}$$

which is (31). □

When $\mathcal{V} = 0$, the covariance identity gives $D_{\text{vis}} = 0$ and $APD\zeta = 0$ almost surely. If $R_{\text{shift}} > 0$, the scalar target removes the complete MSE at fixed radius and $\mathcal{E}(\widetilde{\ell}_\zeta) = 1$. If $R_{\text{shift}} = 0$, the baseline MSE is zero and $\mathcal{E}(\widetilde{\ell}_\zeta)$ is undefined. The ratio Λ is undefined in both cases.

Proof of Proposition 6. Let $X = \ell_\zeta$, $Y = \widetilde{\ell}_\zeta$, $z_T = \ell_T$, and $z_0 = \ell_0$. Since $APh \neq 0$, the vector $\ell_0 - c = APh/r_0$ is nonzero. The actual normalized output is bounded because $\varepsilon > 0$, and the surrogate has a finite second moment by the sub-Gaussian assumption. Indeed, a sub-Gaussian scalar has finite moments of every order, so its fixed finite-dimensional linear images are square-integrable. Thus all expectations below are finite. Expanding $\mathcal{Q}(a)$ gives a one-variable quadratic with leading coefficient $\|\ell_0 - c\|_2^2 > 0$. Finite second moments justify differentiating this polynomial under the expectation, and the resulting first-order condition identifies its unique minimizer as

$$a_{\text{true}} = \frac{\langle \ell_0 - c, \mathbb{E}[X - c]\rangle}{\|\ell_0 - c\|_2^2}.$$

The surrogate at fixed radius is centered around ℓ_T, hence $\mathbb{E}[Y - c] = (\ell_0 - c)/T$. It follows that

$$\begin{aligned}\left|a_{\text{true}} - T^{-1}\right| &= \frac{|\langle \ell_0 - c, \mathbb{E}[X - Y]\rangle|}{\|\ell_0 - c\|_2^2} \\ &\leq \frac{\|\mathbb{E}[X - Y]\|_2}{\|\ell_0 - c\|_2} \leq \frac{(\mathbb{E}\,\|X - Y\|_2^2)^{1/2}}{\|\ell_0 - c\|_2} = \frac{\epsilon_{\text{fr}}}{\|\ell_0 - c\|_2}.\end{aligned}$$

The first inequality is Cauchy–Schwarz in output space. For the second, Jensen's inequality gives $\|\mathbb{E}Z\|_2 \leq \mathbb{E}\,\|Z\|_2$, and Cauchy–Schwarz on the underlying probability space gives $\mathbb{E}\,\|Z\|_2 \leq (\mathbb{E}\,\|Z\|_2^2)^{1/2}$. Square integrability is the relevant hypothesis. These inequalities turn the L^2 perturbation into a bound on the scalar coefficient. The first-order condition for the quadratic risk also yields, for every scalar u,

$$\mathcal{Q}(u) = \mathcal{Q}(a_{\text{true}}) + (u - a_{\text{true}})^2 \|\ell_0 - c\|_2^2.$$

Taking $u = T^{-1}$ and using the preceding bound gives

$$\mathcal{Q}(T^{-1}) - \min_a \mathcal{Q}(a) = (T^{-1} - a_{\text{true}})^2 \|\ell_0 - c\|_2^2 \leq \epsilon_{\text{fr}}^2,$$

which proves both statements in (32).

It remains to compare the relative MSE reductions. Assume now that $\mathcal{V} > 0$, which implies $D_{\text{vis}} > 0$. Define

$$E_T = \mathbb{E}\,\|X - z_T\|_2^2, \qquad E_T^{\text{fr}} = \mathbb{E}\,\|Y - z_T\|_2^2 = D_{\text{vis}},$$
$$E_0 = \mathbb{E}\,\|X - z_0\|_2^2, \qquad E_0^{\text{fr}} = \mathbb{E}\,\|Y - z_0\|_2^2 = R_{\text{shift}} + D_{\text{vis}}.$$

In particular, $0 < E_T^{\mathrm{fr}} \le E_0^{\mathrm{fr}}$. Recall that $\epsilon_{\mathrm{fr}} = (\mathbb{E}\,\|X-Y\|_2^2)^{1/2}$ and $\eta_{\mathrm{fr}} = \epsilon_{\mathrm{fr}}/\sqrt{E_0^{\mathrm{fr}}}$. For any deterministic vector z, expanding $X - z = (Y - z) + (X - Y)$ and applying Cauchy–Schwarz gives

$$\left|\mathbb{E}\,\|X-z\|_2^2 - \mathbb{E}\,\|Y-z\|_2^2\right| \le 2\epsilon_{\mathrm{fr}}\left(\mathbb{E}\,\|Y-z\|_2^2\right)^{1/2} + \epsilon_{\mathrm{fr}}^2.$$

All expectations are finite by the square integrability established above. Applying this inequality with $z = z_T$ and $z = z_0$, and using $E_T^{\mathrm{fr}} \le E_0^{\mathrm{fr}}$, yields

$$|E_T - E_T^{\mathrm{fr}}| \le (2\eta_{\mathrm{fr}} + \eta_{\mathrm{fr}}^2)E_0^{\mathrm{fr}}, \tag{45}$$

$$|E_0 - E_0^{\mathrm{fr}}| \le (2\eta_{\mathrm{fr}} + \eta_{\mathrm{fr}}^2)E_0^{\mathrm{fr}}. \tag{46}$$

Set $a_\eta := 2\eta_{\mathrm{fr}} + \eta_{\mathrm{fr}}^2$. If $\eta_{\mathrm{fr}} \le 1/4$, then $a_\eta < 1$ and (46) implies $E_0 \ge (1 - a_\eta)E_0^{\mathrm{fr}} > 0$. Therefore

$$\begin{aligned}\left|\frac{E_T}{E_0} - \frac{E_T^{\mathrm{fr}}}{E_0^{\mathrm{fr}}}\right| &\le \frac{|E_T - E_T^{\mathrm{fr}}|}{E_0} + \frac{E_T^{\mathrm{fr}}|E_0 - E_0^{\mathrm{fr}}|}{E_0 E_0^{\mathrm{fr}}} \\ &\le \frac{2a_\eta}{1-a_\eta} \le 12\eta_{\mathrm{fr}}.\end{aligned}$$

The last elementary inequality holds for every $0 \le \eta_{\mathrm{fr}} \le 1/4$. Finally, the two relative MSE reductions are one minus the corresponding ratios, and Corollary 5 identifies $1 - E_T^{\mathrm{fr}}/E_0^{\mathrm{fr}} = \Lambda/(1+\Lambda)$. This proves (33). □

B.4 An explicit bound for the error in the predicted MSE reduction

Proposition 6 expresses the law error through the mean squared distance between the actual logits and those at fixed radius. The next corollary turns that distance into a finite-dimensional formula by combining the high-probability event with a uniform bound on its complement.

Assume $|\zeta_i| \le b$ almost surely and $\mathfrak{R}_t \le \theta\bar{r}^2$ for some $\theta \in]0, 1[$. We bound the distance between the output with a random radius and its surrogate at fixed radius in two ways. The quantity g_t bounds this distance on the joint concentration event, while B_b bounds it for every mask. Define

$$g_t := \frac{\chi_t(\|APh\|_2 + \mathfrak{D}_t)}{\bar{r}\sqrt{1-\theta}(1+\sqrt{1-\theta})}, \qquad B_b := \|A\|_{\mathrm{op}} + \frac{\|APh\|_2 + b\sqrt{p}\|APD\|_{\mathrm{op}}}{\bar{r}}, \tag{47}$$

$$\bar{\epsilon}_t := \{g_t^2 + 6e^{-t}B_b^2\}^{1/2}.$$

The bound $\bar{\epsilon}_t$ combines the error on the concentration event with the contribution of its complement, whose probability is at most $6e^{-t}$.

Corollary 12 (An explicit finite-dimensional bound)**.** *Assume $APh \neq 0$. Under these assumptions, $\epsilon_{\mathrm{fr}} \le \bar{\epsilon}_t$. Hence*

$$|a_{\mathrm{true}} - T^{-1}| \le \bar{\epsilon}_t/\|\ell_0 - c\|_2, \qquad \mathcal{Q}(T^{-1}) - \min_a \mathcal{Q}(a) \le \bar{\epsilon}_t^2. \tag{48}$$

If, in addition, $\mathcal{V} > 0$ and $\bar{\epsilon}_t \le \frac{1}{4}\sqrt{R_{\mathrm{shift}} + D_{\mathrm{vis}}}$, then the error in the predicted MSE reduction is at most $12\bar{\epsilon}_t/\sqrt{R_{\mathrm{shift}} + D_{\mathrm{vis}}}$. For inverted Bernoulli dropout, $b = \max\{1, \rho\}$.

Proof of Corollary 12. Let $Z = \|\ell_\zeta - \widetilde{\ell}_\zeta\|_2$ and let $\mathcal{G}_t$ be the joint event in Theorem 2. Its complement has probability at most $6e^{-t}$. On $\mathcal{G}_t$,

$$\begin{aligned}Z &= \|APh + APD\zeta\|_2\left|\frac{1}{r_\zeta} - \frac{1}{\bar{r}}\right| \\ &\le (\|APh\|_2 + \mathfrak{D}_t)\frac{|r_\zeta^2 - \bar{r}^2|}{r_\zeta\bar{r}(r_\zeta + \bar{r})} \le g_t,\end{aligned}$$

where we used $\|APD\zeta\|_2 \le \mathfrak{D}_t$, $|r_\zeta^2 - \bar{r}^2| \le \chi_t\bar{r}^2$, and $r_\zeta \ge \sqrt{1-\theta}\,\bar{r}$ in the last step.

We next control the error away from the good event. For every mask realization, the projector P is a contraction. Indeed, an orthogonal projector has eigenvalues zero or one and therefore satisfies

$\|Pv\|_2 \le \|v\|_2$. Since $r_\zeta^2 = \|P(h + D\zeta)\|_2^2 + \varepsilon$, we obtain

$$\|\ell_\zeta - c\|_2 = \frac{\|AP(h + D\zeta)\|_2}{r_\zeta} \le \|A\|_{\mathrm{op}},$$
$$\|\widetilde{\ell}_\zeta - c\|_2 \le \frac{\|APh\|_2 + \|APD\|_{\mathrm{op}}\|\zeta\|_2}{\bar{r}} \le \frac{\|APh\|_2 + b\sqrt{p}\|APD\|_{\mathrm{op}}}{\bar{r}}.$$

The first line uses $AP = A$, while the last line uses $\|\zeta\|_2 \le b\sqrt{p}$. The triangle inequality therefore gives $Z \le B_b$ everywhere. Splitting the expectation over $\mathcal{G}_t$ and its complement gives

$$\epsilon_{\mathrm{fr}}^2 = \mathbb{E}Z^2 \le g_t^2 + B_b^2 \mathbb{P}(\mathcal{G}_t^c) \le g_t^2 + 6e^{-t}B_b^2 = \bar{\epsilon}_t^2.$$

This split is the identity $\mathbb{E}Z^2 = \mathbb{E}[Z^2 \mathbf{1}_{\{\mathcal{G}_t\}}] + \mathbb{E}[Z^2 \mathbf{1}_{\{\mathcal{G}_t^c\}}]$. It uses only the pointwise bounds on the two events and does not require independence from Z. Taking square roots and substituting the resulting estimate $\epsilon_{\mathrm{fr}} \le \bar{\epsilon}_t$ into Proposition 6 proves all stated bounds. For inverted Bernoulli dropout with keep probability q, $\zeta_i \in \{-1, (1-q)/q\} = \{-1, \rho\}$, so $|\zeta_i| \le \max\{1, \rho\}$. □

B.5 Proofs for replication preserving the original function

For completeness, we first state the two exact identities summarized in the main text. For $k \ge 1$, let $L_k \colon \mathbb{R}^p \longrightarrow \mathbb{R}^{kp}$ replace each v_i by k copies of $v_i/\sqrt{k}$. Index the lifted coordinates by pairs (i, j), where $i \in [\![1, p]\!]$ and $j \in [\![1, k]\!]$. Write e_i for the ith standard basis vector of $\mathbb{R}^p$ and $e_{i,j}$ for the corresponding basis vector of $\mathbb{R}^{kp}$. Then

$$L_k e_i = \frac{1}{\sqrt{k}} \sum_{j=1}^{k} e_{i,j}, \qquad L_k^\top e_{i,j} = \frac{1}{\sqrt{k}} e_i, \qquad L_k^\top L_k = I_p.$$

Set $n = kp$, $h_k = L_k h$, $u_k = L_k u$, $D_k = \mathrm{diag}(u_k)$, and $A_k = AL_k^\top$. Then

$$P_{kp} L_k = L_k P_p \tag{49}$$

for RMS normalization and LayerNorm.

Proposition 13 (The output without dropout remains unchanged)**.** *For every h and k,*

$$c + A_k \mathcal{N}_{P_{kp},\varepsilon}(L_k h) = c + A\mathcal{N}_{P_p,\varepsilon}(h). \tag{50}$$

Proposition 14 (A shared mask gives the same random output)**.** *If $\zeta_{i,j} = \zeta_i$, then, for every mask,*

$$A_k \mathcal{N}_{P_{kp},\varepsilon}(L_k h + \mathrm{diag}(L_k u)\zeta^{\mathrm{shared}}) = A\mathcal{N}_{P_p,\varepsilon}(h + D\zeta). \tag{51}$$

The proof now has three parts. We show that replication commutes with normalization, compare shared and independent masks, and apply the concentration bounds to the independently masked copies.

Let $L = L_k$. For independent copies, write $\zeta^{(k)} = (\zeta_{i,j})_{i,j}$ and define

$$\ell_{k,\zeta} = c + A_k \mathcal{N}_{P_n,\varepsilon}(h_k + D_k \zeta^{(k)}).$$

The corresponding deterministic quantities are

$$r_{0,k}^2 = \|P_n h_k\|_2^2 + \varepsilon, \qquad \bar{r}_k^2 = r_{0,k}^2 + \rho\,\mathrm{Tr}(D_k P_n D_k), \qquad T_k = \frac{\bar{r}_k}{r_{0,k}},$$
$$\ell_{T_k} = c + \frac{A_k P_n h_k}{\bar{r}_k}.$$

For RMS normalization, (49) is immediate. For LayerNorm, if $\bar{v} = p^{-1}\mathbf{1}_p^\top v$, then the mean of Lv is $\bar{v}/\sqrt{k}$. Hence

$$P_n L v = Lv - \frac{\bar{v}}{\sqrt{k}}\mathbf{1}_n = L(v - \bar{v}\mathbf{1}_p) = LP_p v.$$

Transposing also gives $L^\top P_n = P_p L^\top$.

Proof of Proposition 13. The identity $L^\top L = I_p$ means that $\langle Lx, Ly\rangle = \langle x, y\rangle$ for all x, y. In particular, L preserves Euclidean norms, which is the isometry property used below. Indeed, for any $v \in \mathbb{R}^p$, (49) gives

$$\begin{aligned}\mathcal{N}_{P_n,\varepsilon}(Lv) &= \frac{P_n L v}{\sqrt{\|P_n L v\|_2^2 + \varepsilon}} \\ &= \frac{L P_p v}{\sqrt{\|L P_p v\|_2^2 + \varepsilon}} = L\frac{P_p v}{\sqrt{\|P_p v\|_2^2 + \varepsilon}} \\ &= L\mathcal{N}_{P_p,\varepsilon}(v),\end{aligned}$$

where the third equality uses $L^\top L = I_p$. Applying $A_k = AL^\top$ yields

$$A_k \mathcal{N}_{P_n,\varepsilon}(Lv) = AL^\top L \mathcal{N}_{P_p,\varepsilon}(v) = A\mathcal{N}_{P_p,\varepsilon}(v).$$

Taking $v = h$ and adding the unchanged fixed term c proves (50). □

Proof of Proposition 14. Let $\zeta_{i,j}^{\text{shared}} = \zeta_i$. For every (i, j),

$$\left(D_k \zeta^{\text{shared}}\right)_{i,j} = \frac{u_i}{\sqrt{k}}\zeta_i = \left(L(D\zeta)\right)_{i,j}.$$

Consequently,

$$h_k + D_k \zeta^{\text{shared}} = L(h + D\zeta). \tag{52}$$

The identity established in the proof of Proposition 13 holds for every vector. Applying it to $v = h + D\zeta$ proves the pointwise identity (51). □

The visible variance must account for mask dependence. Let $B_k = A_k P_n D_k$ and let Σ_k be the covariance of a centered lifted mask vector. When $A_k P_n h_k \neq 0$, define the ratio accounting for mask covariance

$$\mathcal{V}_{\Sigma,k} := \frac{\operatorname{Tr}(B_k \Sigma_k B_k^\top)}{\|A_k P_n h_k\|_2^2}. \tag{53}$$

For independent copies, $\Sigma_k = \rho I_{kp}$ and this ratio is exactly $\mathcal{V}_k$. For shared copies, the covariance within each group of k replicas is $\rho \mathbf{1}_k \mathbf{1}_k^\top$, and the pointwise identity above gives $\mathcal{V}_{\Sigma,k} = \mathcal{V}_1$. The same identity preserves the random radius and its reference value, so the shared-mask comparison uses the original coefficient $1/T_1$. The independent-coordinate expression $\rho \|B_k\|_{\mathrm{F}}^2 / \|A_k P_n h_k\|_2^2$ still scales as $1/k$, but it does not represent the variance of shared masks.

To assess whether numerator variation is spread across many coordinates or dominated by a few, set $a_i := u_i^2 \|Ae_i\|_2^2$. When $AD \neq 0$, define

$$d_{\text{eff}} := \frac{(\sum_i a_i)^2}{\sum_i a_i^2}, \qquad \mu := \frac{\max_i a_i}{\sum_i a_i}.$$

The quantity d_{eff} is the effective number of contributing coordinates and μ is the largest relative contribution. When $AD = 0$, all a_i vanish and these two ratios are left undefined. The replication identities in the nonzero case are

$$d_{\text{eff},k} = k d_{\text{eff},1}, \qquad \mu_k = \mu_1/k, \qquad d_{\text{eff}}^{-1} \leq \mu \leq d_{\text{eff}}^{-1/2}. \tag{54}$$

Proof of (34) and (54). Let $B_k = A_k P_n D_k$. Using $L^\top P_n = P_p L^\top$ and the canonical assumption $A = AP_p$,

$$B_k = AL^\top D_k. \tag{55}$$

For every lifted coordinate (i, j),

$$B_k e_{i,j} = AL^\top \left(\frac{u_i}{\sqrt{k}} e_{i,j}\right) = \frac{u_i}{k} Ae_i.$$

Consequently,

$$B_k B_k^\top = \sum_{i=1}^{p}\sum_{j=1}^{k} \frac{u_i^2}{k^2}(Ae_i)(Ae_i)^\top = \frac{1}{k}(AD)(AD)^\top.$$

Taking traces gives

$$\|A_k P_n D_k\|_{\mathrm{F}}^2 = \frac{1}{k}\|AP_p D\|_{\mathrm{F}}^2. \tag{56}$$

For any matrix B, the singular-value identities $\|B\|_{\mathrm{F}}^2 = \mathrm{Tr}(BB^\top)$ and $\|B\|_{\mathrm{op}}^2 = \left\|BB^\top\right\|_{\mathrm{op}}$ turn this matrix identity into the Frobenius- and operator-norm scalings used here. The deterministic output direction also satisfies

$$A_k P_n h_k = AL^\top P_n L h = AP_p h.$$

Under the standing condition $AP_p h \neq 0$, substituting these two identities in the definition of visibility proves $\mathcal{V}_k = \mathcal{V}_1/k$. The covariance identity above also gives $\|B_k\|_{\mathrm{op}}^2 = k^{-1}\|AD\|_{\mathrm{op}}^2$. Thus $r_{\mathrm{vis},k} = r_{\mathrm{vis},1}$ when $AD \neq 0$. If $AD = 0$, then $B_k = 0$ for every k, $\mathcal{V}_k = 0$, and the convention $r_{\mathrm{vis},k} = +\infty$ applies.

For later use, we also compute the reference radius. The commutation identity and the isometry of L give

$$r_{0,k}^2 = \|P_n h_k\|_2^2 + \varepsilon = \|P_p h\|_2^2 + \varepsilon = r_0^2.$$

For RMS normalization, $P_n = I_n$ and $\mathrm{Tr}(D_k^2) = \|u_k\|_2^2 = \|u\|_2^2$. For LayerNorm,

$$D_k P_n D_k = D_k^2 - \frac{1}{kp}(Lu)(Lu)^\top,$$

so

$$\mathrm{Tr}(D_k P_n D_k) = \left(1 - \frac{1}{kp}\right)\|u\|_2^2.$$

Therefore

$$\bar{r}_k^2 = \begin{cases} r_0^2 + \rho\|u\|_2^2, & \text{for RMS normalization,} \\ r_0^2 + \rho\left(1 - \dfrac{1}{kp}\right)\|u\|_2^2, & \text{for LayerNorm.} \end{cases} \tag{57}$$

In particular, $T_k = T_1$ for RMS normalization. For LayerNorm, if

$$\bar{r}_\infty^2 = r_0^2 + \rho\|u\|_2^2, \qquad T_\infty = \bar{r}_\infty / r_0,$$

then

$$T_k^2 = T_\infty^2 - \frac{\rho\|u\|_2^2}{kp\, r_0^2}, \qquad T_k = T_\infty + O(k^{-1}).$$

The second relation follows from the mean-value theorem for the square-root map, whose derivative is bounded in a neighborhood of the positive limit T_∞^2. Thus an $O(k^{-1})$ change in the squared ratio produces an $O(k^{-1})$ change in the ratio itself. This identifies the $O(k^{-1})$ LayerNorm centering correction relative to the limiting radius.

Finally, suppose $AD \neq 0$, so the coordinatewise diagnostics are defined. Each contribution a_i is replaced by k contributions a_i/k^2. Their sum is $(\sum_i a_i)/k$, their sum of squares is $(\sum_i a_i^2)/k^3$, and their maximum is $(\max_i a_i)/k^2$. Substitution in the definitions gives $d_{\mathrm{eff},k} = k d_{\mathrm{eff},1}$ and $\mu_k = \mu_1/k$. Finally,

$$\sum_i a_i^2 \leq (\max_i a_i)\sum_i a_i \quad \text{and} \quad (\max_i a_i)^2 \leq \sum_i a_i^2$$

give $d_{\mathrm{eff}}^{-1} \leq \mu \leq d_{\mathrm{eff}}^{-1/2}$. This proves (54) without assigning values to the undefined ratios in the case $AD = 0$. □

Proof of Theorem 7. Fix $t \in [1, +\infty[$. Constants below may depend on the fixed base problem and on t, but they do not depend on k. Replication leaves the mask coordinates independent, centered, and standardized K-sub-Gaussian. Therefore the three estimates in Lemma 10 apply for every k with the same distributional constants. First take $P_p = I_p$. The entries of h_k and u_k are $h_i/\sqrt{k}$ and $u_i/\sqrt{k}$, so

$$\begin{aligned} \|D_k h_k\|_2 &= k^{-1/2}\,\|Dh\|_2\,, \\ \left\|D_k^2\right\|_{\mathrm{F}} &= k^{-1/2}\,\left\|D^2\right\|_{\mathrm{F}}\,, \\ \left\|D_k^2\right\|_{\mathrm{op}} &= k^{-1}\,\left\|D^2\right\|_{\mathrm{op}}\,. \end{aligned}$$

Substitution into (16) gives

$$\mathfrak{R}_{k,t} \le C\Bigg[K\sqrt{\rho}\,\|Dh\|_2\sqrt{\frac{t}{k}} + K^2\rho\|D^2\|_{\mathrm{F}}\sqrt{\frac{t}{k}} + K^2\rho\|D^2\|_{\mathrm{op}}\frac{t}{k}\Bigg]. \tag{58}$$

Equation (57) shows that $\bar{r}_k = \bar{r}_1$ in this case. Therefore $\chi_{t,k} = \mathfrak{R}_{k,t}/\bar{r}_k^2 = O(k^{-1/2})$.

For LayerNorm, let $P_n = I_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top$. Since $P_n h_k = LP_p h$,

$$\|D_k P_n h_k\|_2 = k^{-1/2}\,\|DP_p h\|_2\,.$$

The quadratic matrix is

$$D_k P_n D_k = D_k^2 - \frac{1}{kp}(Lu)(Lu)^\top.$$

For a rank-one matrix $xy^\top$, both its Frobenius norm and its operator norm equal $\|x\|_2\,\|y\|_2$. This standard identity supplies the size of the second term, while the triangle inequality combines it with D_k^2 to give

$$\|D_k P_n D_k\|_{\mathrm{F}} \le \frac{\|D^2\|_{\mathrm{F}}}{\sqrt{k}} + \frac{\|u\|_2^2}{kp},$$
$$\|D_k P_n D_k\|_{\mathrm{op}} \le \frac{\|D^2\|_{\mathrm{op}}}{k} + \frac{\|u\|_2^2}{kp}.$$

Thus $\mathfrak{R}_{k,t} = O(k^{-1/2})$ for LayerNorm as well. Since $\bar{r}_k^2 \ge r_{0,k}^2 \ge \varepsilon$, this also gives $\chi_{t,k} = O(k^{-1/2})$. Equation (57) supplies the additional $O(k^{-1})$ centering correction.

Recall that $B_k = A_k P_n D_k$. The preceding proof established

$$B_k B_k^\top = \frac{1}{k}(AD)(AD)^\top. \tag{59}$$

Consequently,

$$\|B_k\|_{\mathrm{F}} = \frac{\|AD\|_{\mathrm{F}}}{\sqrt{k}}, \qquad \|B_k\|_{\mathrm{op}} = \frac{\|AD\|_{\mathrm{op}}}{\sqrt{k}},$$

with both sides equal to zero when $AD = 0$. The directional threshold in (17) therefore satisfies

$$\mathfrak{D}_{k,t} = CK^2\sqrt{\rho}\left(\|B_k\|_{\mathrm{F}} + \|B_k\|_{\mathrm{op}}\sqrt{t}\right) = O(k^{-1/2}).$$

This also proves $\mathcal{V}_k = \mathcal{V}_1/k$ whenever visibility is defined and shows that $r_{\mathrm{vis},k} = r_{\mathrm{vis},1}$ when $AD \neq 0$.

It remains to establish the absolute output rate in (35). Put

$$v_h = A_k P_n h_k = AP_p h, \qquad r_{k,\zeta} = r(h_k + D_k\zeta^{(k)}).$$

The exact difference between the random output $\ell_{k,\zeta}$ and the deterministic scalar target ℓ_{T_k} is

$$\ell_{k,\zeta} - \ell_{T_k} = v_h\left(\frac{1}{r_{k,\zeta}} - \frac{1}{\bar{r}_k}\right) + \frac{B_k\zeta^{(k)}}{r_{k,\zeta}}.$$

For all sufficiently large k, we have $\chi_{t,k} \le 1/2$. On the joint event of Theorem 2,

$$r_{k,\zeta} \ge \frac{\bar{r}_k}{\sqrt{2}}$$

and the reciprocal-radius identity gives

$$\left|\frac{1}{r_{k,\zeta}} - \frac{1}{\bar{r}_k}\right| \le \frac{\mathfrak{R}_{k,t}}{\bar{r}_k^3\sqrt{1/2}\,(1+\sqrt{1/2})}.$$

Together with $\left\|B_k\zeta^{(k)}\right\|_2 \le \mathfrak{D}_{k,t}$ and $\bar{r}_k \ge \sqrt{\varepsilon}$, these estimates yield

$$\|\ell_{k,\zeta} - \ell_{T_k}\|_2 \le \frac{C_t}{\sqrt{k}}$$

on an event of probability at least $1 - 6e^{-t}$. Given any $\delta \in]0,1[$, choose $t = \log(6/\delta)$, which is fixed with respect to k. By definition, a random error is $O_{\mathbb{P}}(k^{-1/2})$ when, at every fixed confidence level, its magnitude is at most a confidence-dependent constant times $k^{-1/2}$ for all large k. The preceding event proves precisely this $O_{\mathbb{P}}(k^{-1/2})$ rate. The direct argument also covers $AP_ph = 0$, because it uses no relative division by $\|AP_ph\|_2$. □

The conclusion at fixed radius does not yet apply to the original RMS-normalized output, whose denominator remains random. We now show that the discrepancy vanishes under independent replication. Assume inverted Bernoulli dropout, $T_1 > 1$, and $\mathcal{V}_1 > 0$. Define $v_h = A_kP_nh_k = Ah$ and the replicated surrogate at fixed radius by

$$\widetilde{\ell}_{k,\zeta} = c + \frac{v_h + B_k\zeta^{(k)}}{\bar{r}_k}.$$

Set

$$X_k = \left\|\ell_{k,\zeta} - \widetilde{\ell}_{k,\zeta}\right\|_2, \qquad \epsilon_{\mathrm{fr},k} = (\mathbb{E}X_k^2)^{1/2}.$$

Theorem 7 establishes $\ell_{k,\zeta} - \ell_{T_k} = O_{\mathbb{P}}(k^{-1/2})$ and therefore $\ell_{k,\zeta} - \ell_{T_k} \xrightarrow{\mathbb{P}} 0$. Also,

$$\widetilde{\ell}_{k,\zeta} - \ell_{T_k} = \frac{B_k\zeta^{(k)}}{\bar{r}_k} \xrightarrow{\mathbb{P}} 0,$$

by the same directional bound. Hence $X_k \xrightarrow{\mathbb{P}} 0$.

For inverted Bernoulli dropout, the masks are uniformly bounded. Moreover, $\|A_k\|_{\mathrm{op}} \le \|A\|_{\mathrm{op}}$, $\|\mathcal{N}_{P_n,\varepsilon}(v)\|_2 \le 1$, and $B_k\zeta^{(k)} = AD\bar{\zeta}$, where $\bar{\zeta}_i = k^{-1}\sum_{j=1}^{k}\zeta_{i,j}$ remains uniformly bounded because p is fixed. There is therefore a deterministic constant C_X such that $0 \le X_k \le C_X$ for every k and every mask. For any $\eta > 0$,

$$\mathbb{E}X_k^2 \le \eta^2 + C_X^2\mathbb{P}(X_k > \eta).$$

Taking the upper limit as $k \to \infty$ and then letting $\eta \downarrow 0$ shows that $\mathbb{E}X_k^2 \to 0$. Thus $\epsilon_{\mathrm{fr},k} \to 0$. This is the bounded-convergence principle for convergence in probability. The deterministic bound makes X_k^2 uniformly integrable, and the displayed split verifies directly that convergence in probability then implies convergence in L^2. The shift term is positive and independent of k, whereas $D_{\mathrm{vis},k} = D_{\mathrm{vis},1}/k$. Hence $R_{\mathrm{shift},k} + D_{\mathrm{vis},k}$ stays bounded away from zero. For $\eta_{\mathrm{fr},k} = \epsilon_{\mathrm{fr},k}/\sqrt{R_{\mathrm{shift},k} + D_{\mathrm{vis},k}}$, it follows that $\eta_{\mathrm{fr},k} \to 0$ as required by Proposition 6. Since $\Lambda_k = k\Lambda_1$, that proposition now yields

$$\left|\mathcal{E}(\ell_{k,\zeta}) - \frac{k\Lambda_1}{1 + k\Lambda_1}\right| \longrightarrow 0, \qquad \mathcal{E}(\ell_{k,\zeta}) \longrightarrow 1.$$

B.6 Proof of the limiting distribution

The preceding argument gives an upper bound of order $k^{-1/2}$. To determine when the error cannot decrease faster, we keep the first-order random term and apply a multivariate central limit theorem to the mask averages.

Theorem 15 (Gaussian limit)**.** *Under the independent-copy construction with RMS normalization, suppose the entries $\zeta_{i,j}$ are independent and identically distributed, with mean zero, variance $\rho > 0$, and finite fourth moment. As $k \to \infty$ with p, h, u, A, and ε fixed,*

$$\sqrt{k}\left(\ell_{k,\zeta} - c - \frac{Ah}{\bar{r}}\right) \xrightarrow{d} \mathcal{N}(0, \Sigma_\infty), \tag{60}$$

where Σ_∞ is given below. When this matrix is nonzero, the error cannot decrease faster than $k^{-1/2}$ in probability.

Corollary 16 (One-pass approximation in the independently replicated model)**.** *Under independent replication with RMS normalization, for inverted Bernoulli dropout with fixed $q \in]0,1[$, $\varepsilon > 0$, and fixed base problem,*

$$\|\mathbb{E}\,\mathrm{softmax}(\ell_{k,\zeta}) - \mathrm{softmax}(c + Ah/\bar{r})\|_1 = O(k^{-1}). \tag{61}$$

The bound concerns the bias of the mean probabilities.

The numerator depends on averages of the masks, while the squared radius also depends on averages of their squares. Their joint fluctuations therefore involve the covariance of $(\zeta_\circ, \zeta_\circ^2)$, where $\zeta_\circ$ denotes one generic mask entry. Define $\kappa_3 := \mathbb{E}\zeta_\circ^3$ and $\kappa_4 := \mathbb{E}\zeta_\circ^4$, and set

$$\Omega := \mathrm{Cov}\begin{pmatrix}\zeta_\circ \\ \zeta_\circ^2\end{pmatrix} = \begin{pmatrix}\rho & \kappa_3 \\ \kappa_3 & \kappa_4 - \rho^2\end{pmatrix}. \tag{62}$$

For RMS normalization, write $v_h := Ah$ and recall that $\bar{r}^2 = \|h\|_2^2 + \varepsilon + \rho\|u\|_2^2$. Define $G_i := [g_{i,1}, g_{i,2}]$, whose columns describe how changes in these two mask moments affect the normalized output:

$$g_{i,1} := \frac{u_i A e_i}{\bar{r}} - \frac{v_h h_i u_i}{\bar{r}^3}, \qquad g_{i,2} := -\frac{v_h u_i^2}{2\bar{r}^3}. \tag{63}$$

Then the covariance in Theorem 15 is $\Sigma_\infty = \sum_i G_i \Omega G_i^\top$.

Proof of Theorem 15. The limit is taken as $k \to \infty$ with p, h, u, A, and ε fixed. For every original coordinate, define

$$\bar{\zeta}_i = \frac{1}{k}\sum_{j=1}^k \zeta_{i,j}, \qquad \overline{\zeta^2}_i = \frac{1}{k}\sum_{j=1}^k \zeta_{i,j}^2. \tag{64}$$

Since $L^\top D_k \zeta^{(k)} = D\bar{\zeta}$, the numerator and squared radius of the lifted normalized logits are exactly

$$N_k = Ah + AD\bar{\zeta}, \tag{65}$$

$$R_k^2 = \|h\|_2^2 + \varepsilon + 2\sum_{i=1}^p h_i u_i \bar{\zeta}_i + \sum_{i=1}^p u_i^2 \overline{\zeta^2}_i. \tag{66}$$

For each i, the pairs $\big(\zeta_{i,j}, \zeta_{i,j}^2 - \rho\big)$ are independent and identically distributed over j, are centered, and have covariance Ω. The finite fourth-moment assumption gives them a finite second moment. The multivariate central limit theorem (van der Vaart, 1998, Chapter 2) states that normalized averages of independent and identically distributed centered vectors with finite covariance converge to the Gaussian law with that covariance. Its hypotheses hold here, so it gives

$$\sqrt{k}\begin{pmatrix}\bar{\zeta}_i \\ \overline{\zeta^2}_i - \rho\end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \Omega). \tag{67}$$

The source coordinates are mutually independent, so these limiting two-dimensional Gaussian blocks are independent over i. This yields a joint central limit theorem for all $2p$ empirical moments.

Let $\odot$ denote the coordinatewise product. Consider the smooth map

$$F(x,y) = c + \frac{v_h + ADx}{\sqrt{\|h\|_2^2 + \varepsilon + 2(h\odot u)^\top x + (u\odot u)^\top y}}. \tag{68}$$

Because $\varepsilon > 0$, this map is continuously differentiable in a neighborhood of $(x_0, y_0) = (0, \rho\mathbf{1}_p)$, where $F(x_0, y_0) = c + v_h/\bar{r}$. Direct differentiation gives

$$\frac{\partial F}{\partial x_i}(x_0, y_0) = \frac{u_i A e_i}{\bar{r}} - \frac{v_h h_i u_i}{\bar{r}^3} = g_{i,1},$$
$$\frac{\partial F}{\partial y_i}(x_0, y_0) = -\frac{v_h u_i^2}{2\bar{r}^3} = g_{i,2}.$$

The multivariate delta method (van der Vaart, 1998, Chapter 3) states that a continuously differentiable map transports a Gaussian first-order limit through its derivative at the limiting point. The positive stabilizer provides the required differentiability here. Applying this method to (67) yields

$$\sqrt{k}\left(F(\bar{\zeta},\overline{\zeta^2}) - c - \frac{Ah}{\bar{r}}\right) \xrightarrow{d} \mathcal{N}\left(0, \sum_i G_i \Omega G_i^\top\right).$$

Equations (65) and (66) show that $F(\bar{\zeta},\overline{\zeta^2}) = \ell_{k,\zeta}$ exactly. This proves (60). Finally, suppose $\Sigma_\infty \neq 0$. If the centered error were $o_{\mathbb{P}}(k^{-1/2})$, then its product by $\sqrt{k}$ would converge to zero in probability and hence in distribution. Convergence in probability always implies convergence in distribution, and a sequence cannot have two different distributional limits. This contradicts the nontrivial Gaussian limit, which is not concentrated at zero, so the error cannot decrease at that faster rate. □

Proof of Corollary 16. We estimate the difference of the mean predictions directly. The empirical mask moments are centered around their expectations, so their first-order contribution vanishes after averaging.

Let $G = \mathrm{softmax} \circ F$, with F defined in (68), and set

$$z_0 = (0, \rho \mathbf{1}_p), \qquad \Delta_k = (\bar{\zeta},\overline{\zeta^2}) - z_0.$$

Then $G(z_0 + \Delta_k) = \mathrm{softmax}(\ell_{k,\zeta})$, $G(z_0) = \mathrm{softmax}(c + Ah/\bar{r})$, and $\mathbb{E}\Delta_k = 0$.

A Bernoulli mask entry takes values -1 and ρ. Each empirical moment pair $(\bar{\zeta}_i, \overline{\zeta^2}_i)$ therefore lies in the fixed segment $\mathrm{conv}\{(-1,1),(\rho,\rho^2)\}$, as does $(0,\rho)$. The product of these segments is compact and convex. Every point (x,y) in this set satisfies $y_i \geq x_i^2$, so the expression under the square root in F obeys

$$\|h\|_2^2 + \varepsilon + 2(h \odot u)^\top x + (u \odot u)^\top y = \varepsilon + \sum_i (h_i + u_i x_i)^2 + \sum_i u_i^2 (y_i - x_i^2) \geq \varepsilon.$$

Thus G is twice continuously differentiable on a neighborhood of this compact set. Its Hessian satisfies $\|\mathrm{d}^2 G(z)[w,w]\|_1 \leq C_G \|w\|_2^2$ there, where C_G depends only on the fixed base problem and q.

Taylor's formula with integral remainder gives

$$\|G(z_0 + \Delta_k) - G(z_0) - \mathrm{d}G(z_0)\Delta_k\|_1 \leq \frac{C_G}{2}\|\Delta_k\|_2^2.$$

Taking expectations cancels the linear term. Since the mask moments are averages of k independent copies,

$$\mathbb{E}\|\Delta_k\|_2^2 = \mathbb{E}\|\bar{\zeta}\|_2^2 + \mathbb{E}\|\overline{\zeta^2} - \rho \mathbf{1}_p\|_2^2 = \frac{p\{\rho + \kappa_4 - \rho^2\}}{k}.$$

Consequently, Jensen's inequality and the remainder bound yield

$$\|\mathbb{E}\ \mathrm{softmax}(\ell_{k,\zeta}) - \mathrm{softmax}(c + Ah/\bar{r})\|_1 \leq \frac{C_G}{2}\mathbb{E}\|\Delta_k\|_2^2 = \frac{C_G p\{\rho + \kappa_4 - \rho^2\}}{2k} = O(k^{-1}),$$

which proves (61). □

B.7 Proofs for basis dependence

The original prediction remains unchanged when the representation and output map receive inverse orthogonal transformations. For LayerNorm the transformation must also preserve the all-ones direction, while for RMS and Euclidean normalization a Haar average over all orthogonal transformations is available.

Proposition 17 (Rotations compatible with LayerNorm). *Let $e = p^{-1/2}\mathbf{1}_p$ and $P = I_p - ee^\top$. Suppose $Q \in O(p)$ satisfies $Qe = e$. Set $h_Q = Qh$, $u_Q = Qu$, and $A_Q = AQ^\top$. Then*

$$A_Q \mathcal{N}_{P,\varepsilon}(h_Q) = A\mathcal{N}_{P,\varepsilon}(h). \tag{69}$$

The value of T in (7) is also unchanged.

Proof. An orthogonal matrix satisfies $Q^\top Q = I_p$ and therefore preserves inner products and Euclidean norms. This is the only orthogonality property needed in the normalization and radius calculations below. Since Q is orthogonal and $Qe = e$, multiplying the latter identity by $Q^\top$ gives $Q^\top e = e$. Hence

$$Qee^\top = ee^\top \qquad \text{and} \qquad ee^\top Q = ee^\top,$$

so that $QP = PQ$ for $P = I_p - ee^\top$. Therefore

$$\begin{aligned}\mathcal{N}_{P,\varepsilon}(Qh) &= \frac{PQh}{\sqrt{\|PQh\|_2^2 + \varepsilon}} \\ &= \frac{QPh}{\sqrt{\|QPh\|_2^2 + \varepsilon}} = Q\mathcal{N}_{P,\varepsilon}(h),\end{aligned}$$

where the last equality uses preservation of the Euclidean norm by Q. With $A_Q = AQ^\top$, this yields

$$A_Q\mathcal{N}_{P,\varepsilon}(h_Q) = AQ^\top Q\mathcal{N}_{P,\varepsilon}(h) = A\mathcal{N}_{P,\varepsilon}(h),$$

which proves (69).

It remains to verify the radius ratio $T = \bar{r}/r_0$. The radius without dropout is invariant because

$$\|Ph_Q\|_2 = \|PQh\|_2 = \|QPh\|_2 = \|Ph\|_2.$$

For any $v \in \mathbb{R}^p$, writing $D_v = \operatorname{diag}(v)$ and using $P_{ii} = 1 - 1/p$ gives

$$\operatorname{Tr}(D_v P D_v) = \operatorname{Tr}(PD_v^2) = \sum_{i=1}^{p} P_{ii} v_i^2 = \left(1 - \frac{1}{p}\right)\|v\|_2^2.$$

The first equality uses cyclicity of the trace, namely $\operatorname{Tr}(XY) = \operatorname{Tr}(YX)$ for products of compatible finite matrices. Its role is to move the two diagonal factors together so that only the diagonal of P remains. Thus, for the rotated dropout component $u_Q = Qu$,

$$\rho \operatorname{Tr}(\operatorname{diag}(u_Q) P \operatorname{diag}(u_Q)) = \rho\left(1 - \frac{1}{p}\right)\|Qu\|_2^2 = \rho\left(1 - \frac{1}{p}\right)\|u\|_2^2 = \rho \operatorname{Tr}(DPD).$$

Both r_0^2 and the dropout contribution $\rho \operatorname{Tr}(DPD)$ are unchanged. Hence $\bar{r}^2 = r_0^2 + \rho \operatorname{Tr}(DPD)$ and $T = \bar{r}/r_0$ are also unchanged. □

For $P = I_p$ and orthogonal Q, set $h_Q = Qh$, $u_Q = Qu$, and $A_Q = AQ^\top$. Then

$$A_Q\mathcal{N}_{I_p,\varepsilon}(h_Q) = A\mathcal{N}_{I_p,\varepsilon}(h), \tag{70}$$

With $q_i = Q^\top e_i$, the corresponding visibility is given by (37).

Proof of Theorem 8. Let $D_Q = \operatorname{diag}(Qu)$. Since $P = I_p$ and Q is orthogonal,

$$\mathcal{N}_{I_p,\varepsilon}(Qh) = Q\mathcal{N}_{I_p,\varepsilon}(h).$$

Multiplying by $A_Q = AQ^\top$ proves (70). Moreover, $\|Qh\|_2 = \|h\|_2$ and

$$\rho \operatorname{Tr}(D_Q^2) = \rho\|Qu\|_2^2 = \rho\|u\|_2^2,$$

so the baseline radius, the reference radius, and T are unchanged.

For $q_i = Q^\top e_i$, the ith column of the rotated visible-noise matrix is

$$A_Q D_Q e_i = AQ^\top e_i\,(e_i^\top Qu) = Aq_i\,(q_i^\top u).$$

Also, $A_Q h_Q = Ah$. Since $Ah \neq 0$, the definition of visibility therefore gives

$$\mathcal{V}(Q) = \frac{\rho}{\|Ah\|_2^2}\sum_{i=1}^{p}(q_i^\top u)^2\|Aq_i\|_2^2,$$

which is (37).

For a vector q uniformly distributed on the unit sphere S^{p-1}, the standard fourth-moment identity (Meckes, 2019, Chapter 2) gives

$$\mathbb{E}[q_i q_j q_k q_l] = \frac{\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}}{p(p+2)}. \tag{71}$$

For symmetric matrices X, Y, summing the four indices in (71) gives

$$\mathbb{E}[(q^\top X q)(q^\top Y q)] = \frac{\operatorname{Tr}(X)\operatorname{Tr}(Y) + 2\operatorname{Tr}(XY)}{p(p+2)}. \tag{72}$$

Take $X = uu^\top$ and $Y = A^\top A$. Every vector $q_i = Q^\top e_i$ has the same uniform marginal distribution by invariance of Haar measure under fixed orthogonal transformations (Meckes, 2019, Chapter 1). The vectors are dependent, but independence is unnecessary because linearity of expectation gives

$$\begin{aligned}\mathbb{E}_Q \sum_{i=1}^{p} (q_i^\top u)^2 \|Aq_i\|_2^2 &= p\,\mathbb{E}[(q^\top uu^\top q)(q^\top A^\top Aq)] \\ &= \frac{\|u\|_2^2 \|A\|_{\mathrm{F}}^2 + 2\|Au\|_2^2}{p+2}.\end{aligned}$$

Multiplying by $\rho/\|Ah\|_2^2$ proves (38). □

Proof of Proposition 9. Let $r_A = \operatorname{rank}(A)$. By the fundamental theorem of linear algebra,

$$\operatorname{row}(A) = \ker(A)^\perp.$$

In finite-dimensional Euclidean space, this identity also gives dimensions r_A and $p - r_A$ for the two orthogonal subspaces. Hence orthonormal bases of the two spaces concatenate into an orthonormal basis of all of $\mathbb{R}^p$. Choose an orthonormal basis $q_1, \ldots, q_{r_A}$ of $\operatorname{row}(A)$ and an orthonormal basis $q_{r_A+1}, \ldots, q_p$ of $\ker(A)$. Together these vectors form an orthonormal basis of $\mathbb{R}^p$. Define $Q \in O(p)$ by $Q^\top e_i = q_i$.

For every $i > r_A$, we have $q_i \in \ker(A)$ and therefore $Aq_i = 0$. Equivalently, the ith column of the rotated output map $A_Q = AQ^\top$ is

$$A_Q e_i = AQ^\top e_i = Aq_i = 0.$$

Thus coordinates with index $i > r_A$ cannot contribute directly to the numerator reaching the output. They may nevertheless change the normalization radius. □

C Further implementation remarks

Root mean square convention. RMSNorm and LayerNorm divide by a root mean square. As explained in Section 3, multiplying the numerator and denominator by $\sqrt{p}$ gives the Euclidean radius convention used in the proofs. Thus $\kappa = \sqrt{p}$ for RMSNorm and LayerNorm, while $\kappa = 1$ for Euclidean normalization, and ε becomes p times the implementation stabilizer in the RMS cases. All ratios of radii, including T, remain unchanged. For the replication L_k in (50), keeping the Euclidean stabilizer ε fixed requires dividing the implementation stabilizer by k for both RMSNorm and LayerNorm.

Learned gains. If a normalization layer outputs $\kappa\Gamma\mathcal{N}_{P,\varepsilon}(v) + \beta_{\mathrm{N}}$ with diagonal Γ, set $A = \kappa C_m W \Gamma P$ and $c = C_m(W\beta_{\mathrm{N}} + b)$. The extra projector does not change the function because $\mathcal{N}_{P,\varepsilon}(v)$ lies in the image of P, and it restores the canonical identity $AP = A$. All statements then apply without change.

Scalar rescaling with a fixed output component. The analytic scalar leaves the fixed vector c unchanged. If the uncentered fixed logit component is class independent, output centering removes it. For softmax outputs with $c = 0$, the map $\ell_0 \mapsto \ell_0/T$ has the algebraic form of temperature scaling. The radius ratio T is computed from the normalization radius. If c is class dependent, the correct deterministic approximation is $c + (\ell_0 - c)/T$.

Zero stabilizer. We assume $\varepsilon > 0$ throughout the main results. If $\varepsilon = 0$, the identities involving ℓ_0 and T remain valid only when $r_0 > 0$, and only for mask realizations satisfying $r_\zeta > 0$. Bounds involving the inverse random radius also require a positive lower bound on r_ζ on the stated probability event. The Euclidean-normalization experiments use a numerical floor on the norm rather than a positive additive stabilizer. The zero-stabilizer normalization agrees with this implementation only where the floor is inactive.

D SEVERAL NONLINEAR LAYERS

The main analysis places the random representation $h + M\zeta$ directly before normalization $\mathcal{N}_{P,\varepsilon}$ and the linear output matrix A. We now move the random change earlier and let it pass through several ReLU layers. The final change is still a linear function of the masks, apart from an explicit term created when ReLU units change state. As before, the input and all trained weights stay fixed while only the masks are random.

D.1 EXTENSION TO DEEPER RELU LAYERS

Model and linear propagation. Consider L residual ReLU blocks with $W_{1,\ell} \in \mathbb{R}^{w_\ell \times p}$, $W_{2,\ell} \in \mathbb{R}^{p \times w_\ell}$, and $b_\ell \in \mathbb{R}^{w_\ell}$:

$$F_\ell(v) = v + W_{2,\ell}\,\mathrm{ReLU}(W_{1,\ell}v + b_\ell), \qquad \ell \in [\![0, L-1]\!], \tag{73}$$

and let $h_{\ell+1} = F_\ell(h_\ell)$ denote the states obtained without any random change. Independent random changes may enter at several layers:

$$\widetilde{h}_0 = h_0 + D_0\zeta_0, \qquad \widetilde{h}_{\ell+1} = F_\ell(\widetilde{h}_\ell) + D_{\ell+1}\zeta_{\ell+1}. \tag{74}$$

Each matrix $D_s \in \mathbb{R}^{p \times d_s}$ is fixed and $\zeta_s \in \mathbb{R}^{d_s}$. The entries of $\zeta = (\zeta_0^\top, \dots, \zeta_L^\top)^\top$ are mutually independent, have mean zero, have variance $\rho > 0$, and satisfy the same assumption as in (11), namely $\|\zeta_j/\sqrt{\rho}\|_{\psi_2} \le K$. Keeping only one nonzero D_s describes dropout added at one earlier layer. Several nonzero matrices describe independent random changes added at several layers.

For each block, define

$$a_\ell = W_{1,\ell}h_\ell + b_\ell, \qquad G_\ell = \mathrm{diag}\big(\mathbf{1}_{\{a_\ell > 0\}}\big), \qquad J_\ell = I_p + W_{2,\ell}G_\ell W_{1,\ell}. \tag{75}$$

The inequality in the indicator is interpreted coordinatewise. The diagonal matrix G_ℓ records which ReLU units are active before the random change. When every preactivation is nonzero, J_ℓ is the Jacobian of block ℓ at this state. In general, it is the selected local linear map associated with the convention $\mathbf{1}_{\{a_{\ell j} > 0\}}$. For $s < t$, write $J_{t:s} = J_{t-1} \cdots J_s$, with $J_{s:s} = I_p$. The matrix that propagates all random sources to layer t is

$$\mathcal{M}_t = \big[J_{t:0}D_0\; J_{t:1}D_1\; \cdots\; J_{t:t}D_t\big], \qquad M_{\mathrm{eff}} \coloneqq \mathcal{M}_L. \tag{76}$$

Thus M_{eff} contains one product of selected local linear maps for each random source. It describes the complete linear effect at the final normalized representation.

Theorem 18 (Exact split into a linear part and ReLU changes)**.** *Let $\Delta_\ell = \widetilde{h}_\ell - h_\ell$ be the change at layer ℓ. Define the part that is missed by the local linear formula as*

$$r_\ell(\Delta) = \mathrm{ReLU}(a_\ell + W_{1,\ell}\Delta) - \mathrm{ReLU}(a_\ell) - G_\ell W_{1,\ell}\Delta. \tag{77}$$

For every realization of the masks,

$$\Delta_L = M_{\mathrm{eff}}\zeta + g_\zeta, \qquad g_\zeta = \sum_{\ell=0}^{L-1} J_{L:\ell+1}W_{2,\ell}r_\ell(\Delta_\ell). \tag{78}$$

If $\mu_\ell = |a_\ell|$ gives the distances of the ReLU inputs from zero, then

$$\|g_\zeta\|_2 \le \sum_{\ell=0}^{L-1} \|J_{L:\ell+1}W_{2,\ell}\|_{\mathrm{op}} \left\| \big(|W_{1,\ell}\Delta_\ell| - \mu_\ell\big)_+ \right\|_2. \tag{79}$$

In particular, $g_\zeta = 0$ when every ReLU unit keeps the same active or inactive state. In that case, the final representation is exactly $h_L + M_{\mathrm{eff}}\zeta$.

Proof. From the definitions of F_ℓ, J_ℓ, and r_ℓ, we have, for every realization of the masks,

$$\begin{aligned}\Delta_{\ell+1} &= F_\ell(h_\ell + \Delta_\ell) - F_\ell(h_\ell) + D_{\ell+1}\zeta_{\ell+1} \\ &= \Delta_\ell + W_{2,\ell}\left[G_\ell W_{1,\ell}\Delta_\ell + r_\ell(\Delta_\ell)\right] + D_{\ell+1}\zeta_{\ell+1} \\ &= J_\ell\Delta_\ell + W_{2,\ell}r_\ell(\Delta_\ell) + D_{\ell+1}\zeta_{\ell+1}.\end{aligned}$$

Write $\zeta_{0:t} = (\zeta_0^\top, \dots, \zeta_t^\top)^\top$. We claim that for every $t \in [\![0, L]\!]$,

$$\Delta_t = \mathcal{M}_t\zeta_{0:t} + \sum_{\ell=0}^{t-1} J_{t:\ell+1}W_{2,\ell}r_\ell(\Delta_\ell). \tag{80}$$

For $t = 0$, the sum is empty and $\Delta_0 = D_0\zeta_0 = \mathcal{M}_0\zeta_0$. Suppose the identity holds at time t. Substitution into the one-step recursion gives the same formula at time $t+1$, because

$$\mathcal{M}_{t+1} = [J_t\mathcal{M}_t \quad D_{t+1}] \qquad \text{and} \qquad J_tJ_{t:\ell+1} = J_{t+1:\ell+1}.$$

Induction proves (80). Taking $t = L$ gives (78).

It remains to bound the local remainders. For scalars a, u, define

$$\varphi(a, u) = (a + u)_+ - a_+ - \mathbf{1}_{\{a>0\}}u.$$

If $\mathbf{1}_{\{a+u>0\}} = \mathbf{1}_{\{a>0\}}$, both points lie in the same affine region of ReLU and $\varphi(a, u) = 0$. If $a > 0$ and $a + u \le 0$, then $u \le -a$ and

$$|\varphi(a, u)| = -a - u = |u| - |a|.$$

If $a \le 0$ and $a + u > 0$, then $u > |a|$ and

$$|\varphi(a, u)| = a + u = |u| - |a|.$$

Thus, in every case,

$$|\varphi(a, u)| \le (|u| - |a|)_+.$$

Applying this inequality coordinatewise with $a = a_\ell$ and $u = W_{1,\ell}\Delta_\ell$ gives

$$\|r_\ell(\Delta_\ell)\|_2 \le \left\|\left(|W_{1,\ell}\Delta_\ell| - \mu_\ell\right)_+\right\|_2.$$

The operator norm is subordinate to the Euclidean norm, which means $\|Bx\|_2 \le \|B\|_{\mathrm{op}}\|x\|_2$ for every deterministic matrix B and vector x. Together with the triangle inequality, this gives

$$\begin{aligned}\|g_\zeta\|_2 &\le \sum_{\ell=0}^{L-1} \|J_{L:\ell+1}W_{2,\ell}\|_{\mathrm{op}}\,\|r_\ell(\Delta_\ell)\|_2 \\ &\le \sum_{\ell=0}^{L-1} \|J_{L:\ell+1}W_{2,\ell}\|_{\mathrm{op}}\left\|\left(|W_{1,\ell}\Delta_\ell| - \mu_\ell\right)_+\right\|_2,\end{aligned}$$

which proves (79). If no ReLU unit changes state, every local remainder is zero. Hence $g_\zeta = 0$ and $\widetilde{h}_L = h_L + M_{\mathrm{eff}}\zeta$ exactly. □

The decomposition in Theorem 18 holds pointwise for every mask. Its linear component is most informative when few ReLU units change state. The sub-Gaussian assumption permits a quantitative bound on the probability of such a change. Define the linear change at layer ℓ by

$$\widehat{\Delta}_\ell = \mathcal{M}_\ell(\zeta_0^\top, \dots, \zeta_\ell^\top)^\top, \qquad \tau_{\ell j}^2 = \rho\left\|w_{\ell j}^\top\mathcal{M}_\ell\right\|_2^2,$$

where $w_{\ell j}^\top$ is row j of $W_{1,\ell}$. The linear preactivation change $w_{\ell j}^\top\widehat{\Delta}_\ell$ has mean zero and standard deviation $\tau_{\ell j}$. Comparing this standard deviation with the margin $|a_{\ell j}|$ quantifies the risk of a state change.

Proposition 19 (Probability that at least one ReLU changes state)**.** *Assume $\mu_{\ell j} = |a_{\ell j}| > 0$ for every ReLU unit. There is an absolute constant $c_0 > 0$ such that*

$$\mathbb{P}(\textit{at least one ReLU changes state}) \le \delta_{\mathrm{gate}} := \min\left\{1, 2\sum_{\ell=0}^{L-1}\sum_{j}\exp\left(-c_0\frac{\mu_{\ell j}^2}{K^2\tau_{\ell j}^2}\right)\right\}, \tag{81}$$

with a zero summand when $\tau_{\ell j} = 0$.

Proof. Let $\mathcal{G}$ be the event that at least one ReLU changes state, and suppose that $\mathcal{G}$ occurs. Let ℓ be the smallest layer containing a changed unit. At every preceding layer, the perturbed and unperturbed preactivations have the same ReLU indicators. Hence all earlier local remainders vanish, and the exact recursion from Theorem 18 gives

$$\Delta_\ell = \mathcal{M}_\ell(\zeta_0^\top, \dots, \zeta_\ell^\top)^\top = \widehat{\Delta}_\ell.$$

If unit j changes state at this layer, then $a_{\ell j}$ and $a_{\ell j} + w_{\ell j}^\top\widehat{\Delta}_\ell$ lie on opposite sides of zero. Since $\mu_{\ell j} = |a_{\ell j}| > 0$, this implies

$$|w_{\ell j}^\top\widehat{\Delta}_\ell| \ge \mu_{\ell j}.$$

Consequently,

$$\mathcal{G} \subseteq \bigcup_{\ell=0}^{L-1}\bigcup_{j}\left\{|w_{\ell j}^\top\widehat{\Delta}_\ell| \ge \mu_{\ell j}\right\}.$$

For a fixed pair (ℓ, j), set $b_{\ell j} = \mathcal{M}_\ell^\top w_{\ell j}$. Then

$$w_{\ell j}^\top\widehat{\Delta}_\ell = b_{\ell j}^\top(\zeta_0^\top, \dots, \zeta_\ell^\top)^\top.$$

The concatenated source coordinates are independent and centered, and their standardized versions have sub-Gaussian norm at most K. When $\tau_{\ell j} > 0$, apply the linear-form bound from Lemma 10 to the standardized source coordinates. Their sub-Gaussian norms are bounded by K. Restoring the factor $\sqrt{\rho}$ gives the two-sided tail bound $2\exp(-cu^2/(K^2\rho\,\|b\|_2^2))$. Taking $b = b_{\ell j}$ gives

$$\mathbb{P}\left(|w_{\ell j}^\top\widehat{\Delta}_\ell| \ge \mu_{\ell j}\right) \le 2\exp\left(-c_0\frac{\mu_{\ell j}^2}{K^2\rho\,\|b_{\ell j}\|_2^2}\right).$$

By definition,

$$\rho\,\|b_{\ell j}\|_2^2 = \rho\left\|w_{\ell j}^\top\mathcal{M}_\ell\right\|_2^2 = \tau_{\ell j}^2.$$

If $\tau_{\ell j} = 0$, the corresponding linear form is identically zero, so the crossing event is impossible because $\mu_{\ell j} > 0$. This justifies the zero-summand convention. The union bound needs no independence between the crossing events and bounds their union by the sum of their probabilities. Applying it over all (ℓ, j) and truncating the resulting upper bound at one proves (81). □

Connection with the main results. Set $h := h_L$ and $M := M_{\mathrm{eff}}$. The linear approximation is exactly the model $h + M\zeta$ from Proposition 11. We use the same notation:

$$\bar{r}_M^2 = r_0^2 + \rho\,\mathrm{Tr}(M^\top PM), \qquad T_M = \frac{\bar{r}_M}{r_0}, \qquad \mathcal{V}_M = \frac{\rho\,\|APM\|_{\mathrm{F}}^2}{\|APh\|_2^2}, \qquad \Lambda_M = \frac{(T_M-1)^2}{\mathcal{V}_M}, \tag{82}$$

with $r_0^2 = \|Ph\|_2^2 + \varepsilon$. The ratio $\mathcal{V}_M$ is defined when $APh \neq 0$, and Λ_M is defined when $\mathcal{V}_M > 0$. To compare the deep output with the model with a fixed radius, we first omit the ReLU remainder and then replace the random radius. The following outputs and root mean squared errors describe these two successive approximations:

$$L_\zeta^{\mathrm{deep}} := \ell(h + M\zeta + g_\zeta), \qquad L_\zeta^{\mathrm{aff}} := \ell(h + M\zeta),$$

$$\epsilon_{\mathrm{gate}} := \left(\mathbb{E}\left\|L_\zeta^{\mathrm{deep}} - L_\zeta^{\mathrm{aff}}\right\|_2^2\right)^{1/2}, \qquad \epsilon_{\mathrm{fr},M} := \left(\mathbb{E}\left\|L_\zeta^{\mathrm{aff}} - \widetilde{\ell}_{\zeta,M}\right\|_2^2\right)^{1/2},$$

where $\widetilde{\ell}_{\zeta,M} = c + (APh + APM\zeta)/\bar{r}_M$ is the model in which the random radius is replaced by $\bar{r}_M$.

For $\varepsilon > 0$, the normalization map is globally $\varepsilon^{-1/2}$-Lipschitz. Indeed, with $r(v) = (\|Pv\|_2^2 + \varepsilon)^{1/2}$,

$$\mathrm{d}\mathcal{N}_{P,\varepsilon}(v) = \left(\frac{I_p}{r(v)} - \frac{(Pv)(Pv)^\top}{r(v)^3}\right) P, \qquad \|\mathrm{d}\mathcal{N}_{P,\varepsilon}(v)\|_{\mathrm{op}} \le \frac{1}{r(v)} \le \frac{1}{\sqrt{\varepsilon}}. \tag{83}$$

Integrating the derivative along the segment between two inputs gives the global Lipschitz bound. This bounds the output error by the propagated ReLU remainder. Under the positive-margin assumptions of Proposition 19, the probability of any ReLU state change gives a second bound. Combining the two bounds yields

$$\epsilon_{\mathrm{gate}} \le \min\left\{ \frac{\|A\|_{\mathrm{op}}}{\sqrt{\varepsilon}} \left(\mathbb{E}\,\|g_\zeta\|_2^2\right)^{1/2}, 2\,\|A\|_{\mathrm{op}} \sqrt{\delta_{\mathrm{gate}}} \right\}. \tag{84}$$

The first term follows from the Lipschitz bound. For the second, the two models are equal when no ReLU changes state. On the complementary event, each normalized representation has norm at most one, so their output difference is at most $2\,\|A\|_{\mathrm{op}}$. The probability of this event is at most δ_{gate}, which gives the stated root mean squared bound. Let $\epsilon_{\mathrm{deep},M} = \epsilon_{\mathrm{gate}} + \epsilon_{\mathrm{fr},M}$ and define, as in (25),

$$R_{\mathrm{shift},M} = \frac{(T_M - 1)^2\,\|APh\|_2^2}{\bar{r}_M^2}, \qquad D_{\mathrm{vis},M} = \frac{\rho\,\|APM\|_{\mathrm{F}}^2}{\bar{r}_M^2}.$$

For any square-integrable random output L whose baseline MSE is positive, the relative MSE reduction obtained by the analytic target based on T_M is

$$\mathcal{E}_M(L) := \frac{\mathbb{E}\|L - \ell_0\|_2^2 - \mathbb{E}\|L - \{c + (\ell_0 - c)/T_M\}\|_2^2}{\mathbb{E}\|L - \ell_0\|_2^2}, \tag{85}$$

The next theorem separates the error caused by ReLU state changes from the error caused by replacing the random radius.

Theorem 20 (Law relating the mean shift to dispersion after several ReLU layers)**.** *Assume $APh \neq 0$ and $\mathcal{V}_M > 0$. Set*

$$\mathcal{Q}_{\mathrm{deep}}(a) = \mathbb{E}\|L_\zeta^{\mathrm{deep}} - \{c + a(\ell_0 - c)\}\|_2^2,$$

and let a_{deep} minimize $\mathcal{Q}_{\mathrm{deep}}$. Then

$$|a_{\mathrm{deep}} - T_M^{-1}| \le \frac{\epsilon_{\mathrm{deep},M}}{\|\ell_0 - c\|_2}, \qquad \mathcal{Q}_{\mathrm{deep}}(T_M^{-1}) - \min_a \mathcal{Q}_{\mathrm{deep}}(a) \le \epsilon_{\mathrm{deep},M}^2. \tag{86}$$

If

$$\eta_M := \frac{\epsilon_{\mathrm{deep},M}}{\sqrt{R_{\mathrm{shift},M} + D_{\mathrm{vis},M}}} \le \frac{1}{4},$$

then

$$\left| \mathcal{E}_M(L_\zeta^{\mathrm{deep}}) - \frac{\Lambda_M}{1 + \Lambda_M} \right| \le 12\eta_M. \tag{87}$$

Proof. For a random vector Z, write $\|Z\|_{L^2} = (\mathbb{E}\,\|Z\|_2^2)^{1/2}$. Set

$$X = L_\zeta^{\mathrm{deep}}, \qquad Z = L_\zeta^{\mathrm{aff}}, \qquad Y = \widetilde{\ell}_{\zeta,M}.$$

The two normalized outputs are bounded, and the surrogate at fixed radius has a finite second moment. Minkowski's inequality therefore gives

$$\|X - Y\|_{L^2} \le \|X - Z\|_{L^2} + \|Z - Y\|_{L^2} = \epsilon_{\mathrm{gate}} + \epsilon_{\mathrm{fr},M} = \epsilon_{\mathrm{deep},M}.$$

Denote $\|X - Y\|_{L^2}$ by δ.

The surrogate at fixed radius can be written as

$$Y = c + \frac{\ell_0 - c}{T_M} + \nu, \qquad \nu = \frac{APM\zeta}{\bar{r}_M}.$$

Because the source coordinates are centered with covariance $\rho I_{\sum_{s=0}^{L} d_s}$,

$$\mathbb{E}\nu = 0, \qquad \mathbb{E}\,\|\nu\|_2^2 = \frac{\rho\,\|APM\|_{\mathrm{F}}^2}{\bar{r}_M^2} = D_{\mathrm{vis},M}.$$

This uses the covariance identity $\mathbb{E}\,\|C\xi\|_2^2 = \mathrm{Tr}(C\,\mathrm{Cov}(\xi)C^\top)$ for a centered finite-variance vector ξ. It turns the expected random squared norm into the Frobenius norm of APM. Therefore, for every scalar a,

$$\mathbb{E}\|Y - \{c + a(\ell_0 - c)\}\|_2^2 = \left(a - T_M^{-1}\right)^2 \|\ell_0 - c\|_2^2 + D_{\mathrm{vis},M}. \tag{88}$$

In particular, the exact surrogate optimizer is T_M^{-1}.

Since $APh \neq 0$, we have $\ell_0 - c \neq 0$. Expanding the quadratic risk of X and differentiating is legitimate because its coefficients have finite moments. The positive leading coefficient $\|\ell_0 - c\|_2^2$ makes the first-order condition the unique minimizer, giving

$$a_{\mathrm{deep}} = \frac{\langle \ell_0 - c, \mathbb{E}[X - c]\rangle}{\|\ell_0 - c\|_2^2}.$$

Using $\mathbb{E}[Y - c] = (\ell_0 - c)/T_M$, Cauchy–Schwarz bounds the deterministic inner product. Jensen's inequality for the convex norm, followed by Cauchy–Schwarz on the underlying probability space, gives $\|\mathbb{E}U\|_2 \le (\mathbb{E}\,\|U\|_2^2)^{1/2}$. These tools apply because $U = X - Y$ is square-integrable, and they yield

$$\begin{aligned} |a_{\mathrm{deep}} - T_M^{-1}| &= \frac{|\langle \ell_0 - c, \mathbb{E}[X - Y]\rangle|}{\|\ell_0 - c\|_2^2} \\ &\le \frac{\|\mathbb{E}[X - Y]\|_2}{\|\ell_0 - c\|_2} \le \frac{\delta}{\|\ell_0 - c\|_2} \le \frac{\epsilon_{\mathrm{deep},M}}{\|\ell_0 - c\|_2}. \end{aligned}$$

The first-order condition for this quadratic also implies

$$\mathcal{Q}_{\mathrm{deep}}(a) = \mathcal{Q}_{\mathrm{deep}}(a_{\mathrm{deep}}) + (a - a_{\mathrm{deep}})^2 \|\ell_0 - c\|_2^2.$$

Taking $a = T_M^{-1}$ and using the preceding coefficient bound yields

$$\mathcal{Q}_{\mathrm{deep}}(T_M^{-1}) - \min_a \mathcal{Q}_{\mathrm{deep}}(a) \le \delta^2 \le \epsilon_{\mathrm{deep},M}^2,$$

which proves (86).

It remains to compare the relative MSE reductions. Define

$$z_T = c + (\ell_0 - c)/T_M, \qquad z_0 = \ell_0,$$

and

$$\begin{aligned} N &= \mathbb{E}\,\|X - z_T\|_2^2, & N_0 &= \mathbb{E}\,\|Y - z_T\|_2^2 = D_{\mathrm{vis},M}, \\ D &= \mathbb{E}\,\|X - z_0\|_2^2, & D_0 &= \mathbb{E}\,\|Y - z_0\|_2^2 = R_{\mathrm{shift},M} + D_{\mathrm{vis},M}. \end{aligned}$$

Equation (88) gives

$$1 - \frac{N_0}{D_0} = \frac{R_{\mathrm{shift},M}}{R_{\mathrm{shift},M} + D_{\mathrm{vis},M}} = \frac{\Lambda_M}{1 + \Lambda_M}.$$

The assumption $\mathcal{V}_M > 0$ ensures $D_{\mathrm{vis},M} > 0$ and hence $D_0 > 0$.

For any deterministic vector z, expanding $X - z = (Y - z) + (X - Y)$ and applying Cauchy–Schwarz gives

$$\left|\mathbb{E}\,\|X - z\|_2^2 - \mathbb{E}\,\|Y - z\|_2^2\right| \le 2\delta \left(\mathbb{E}\,\|Y - z\|_2^2\right)^{1/2} + \delta^2.$$

Cauchy–Schwarz is applied in L^2. Finite second moments are its relevant hypothesis, and it controls the expected cross term between $Y - z$ and $X - Y$. Let $\gamma = \delta/\sqrt{D_0}$. Since $N_0 \le D_0$, the preceding bound with $z = z_T$ and $z = z_0$ yields

$$|N - N_0| \le (2\gamma + \gamma^2)D_0, \qquad |D - D_0| \le (2\gamma + \gamma^2)D_0.$$

Moreover,

$$\gamma \le \frac{\epsilon_{\mathrm{deep},M}}{\sqrt{D_0}} = \eta_M .$$

If $\eta_M \le 1/4$, then $\gamma \le 1/4$. Put $b_\gamma = 2\gamma + \gamma^2 < 1$. The denominator bound implies $D \ge (1 - b_\gamma)D_0 > 0$, and therefore

$$\begin{aligned}\left|\frac{N}{D} - \frac{N_0}{D_0}\right| &\le \frac{|N - N_0|}{D} + \frac{N_0|D - D_0|}{DD_0} \\ &\le \frac{2b_\gamma}{1 - b_\gamma} \le 12\gamma \le 12\eta_M .\end{aligned}$$

The actual and reference relative MSE reductions are one minus these two ratios. Combining the last inequality with the identity at fixed radius proves (87). □

The block columns in M_{eff} correspond to the different random sources, so their squared Frobenius norms add algebraically:

$$\rho \operatorname{Tr}(M_{\mathrm{eff}}^\top P M_{\mathrm{eff}}) = \rho \sum_{s=0}^{L} \|P J_{L:s} D_s\|_{\mathrm{F}}^2 , \qquad \rho \|A P M_{\mathrm{eff}}\|_{\mathrm{F}}^2 = \rho \sum_{s=0}^{L} \|A P J_{L:s} D_s\|_{\mathrm{F}}^2 . \tag{89}$$

Independence of the source coordinates then makes the same sums their total variance contributions, without cross-covariance terms. Depth alone therefore does not determine the result. The relevant quantities are the expected change in the squared radius after propagation through the later layers and the part of the random change that reaches the output through A.

Dropout configurations covered by the deep extension. The recursion (74) covers dropout at one earlier layer and independent random changes built from branch values fixed before the masks are drawn. With standard dropout at several layers, a later mask multiplies a branch whose input already depends on earlier masks, adding another nonlinear error term. The stability theorem still applies when this extra term is small at the logits, while the explicit probability δ_{gate} concerns the independent additive model above.

The same argument applies to a continuously differentiable map $F\colon \mathbb{R}^p \to \mathbb{R}^p$. Let $D \in \mathbb{R}^{p\times d}$, $\zeta \in \mathbb{R}^d$, and $M = \mathrm{d}F(h)D \in \mathbb{R}^{p\times d}$. Assume that $\mathrm{d}F$ is H_F-Lipschitz in operator norm on a fixed set containing every segment $[h, h + D\zeta]$, so that

$$\|\mathrm{d}F(v) - \mathrm{d}F(w)\|_{\mathrm{op}} \le H_F \|v - w\|_2$$

on this set. Taylor's formula gives

$$\|F(h + D\zeta) - F(h) - M\zeta\|_2 \le \frac{H_F}{2}\|D\zeta\|_2^2 .$$

Using (83), the resulting root mean squared output error satisfies

$$\left(\mathbb{E}\,\|A\,[\mathcal{N}_{P,\varepsilon}(F(h + D\zeta)) - \mathcal{N}_{P,\varepsilon}(F(h) + M\zeta)]\|_2^2\right)^{1/2} \le \frac{\|A\|_{\mathrm{op}} H_F}{2\sqrt{\varepsilon}} \left(\mathbb{E}\,\|D\zeta\|_2^4\right)^{1/2} . \tag{90}$$

The sub-Gaussian assumption gives the required fourth moment. For this smooth-map model, the displayed bound controls the nonlinear output error. Adding the radius-replacement error $\epsilon_{\mathrm{fr},M}$ then bounds the root mean squared distance to the surrogate at fixed radius, with base representation $F(h)$ and $M = \mathrm{d}F(h)D$.

D.2 Experiments with several ReLU layers

Protocol. We train three residual ReLU MLPs on 50 000 MNIST training images, using 10 000 disjoint validation images and a maximum of seven epochs. We retain the checkpoint with the smallest validation cross-entropy, then fix its weights. Each model has a 96-dimensional encoder, eight residual blocks with hidden dimension 192, residual scale 0.25, Euclidean feature normalization, and a centered cosine output. For each seed, we evaluate 512 test images with 160 independent masks and $q \in \{0.40, 0.55, 0.70, 0.85, 0.95\}$. The parameter q is the probability of keeping a coordinate, and

we analyze every input separately without labels. Dropout is absent during training and is introduced only for these interventions. A common scalar is fitted on 128 additional masks, independent of the 160 evaluation masks. Training and checkpoint selection follow Appendix E.2.

The three settings separate different sources of difficulty. In *one earlier source*, dropout is inserted before the last $k \in \{0, 1, 2, 4, 8\}$ blocks, which varies the number of later blocks without changing the trained classifier. In *fixed branch outputs*, every block receives an independent random change computed before any mask is drawn, so (74) holds exactly. In *standard layerwise dropout*, each mask multiplies a branch evaluated at an already perturbed state, which creates the interaction described above. The complete experiment has 105 groups defined by setting, seed, q, and the number of later blocks. Each group contains 512 input-condition records, and each record aggregates 160 masks, for $53\,760$ records in total.

For every setting, define the total residual

$$g_\zeta^{\mathrm{tot}} = \Delta_L - M_{\mathrm{eff}}\zeta.$$

In the one-source and fixed-branch settings, the additive recursion applies and $g_\zeta^{\mathrm{tot}} = g_\zeta$ from Theorem 18. Under standard layerwise dropout, g_ζ^{tot} also contains the interactions created when a later mask multiplies an already perturbed branch.

For every input, we compute all J_ℓ and M_{eff} directly from the network, without fitting a linear map. We report the observed relative MSE reduction $\mathcal{E}_{\mathrm{obs}}$, the prediction $\mathcal{E}_{\mathrm{pred}} = \Lambda_M/(1+\Lambda_M)$, and the common scalar $\widehat{a}$ fitted on the 128 fitting masks. We also compute

$$\widehat{G} = \frac{\left(N_{\mathrm{mask}}^{-1}\sum_r \left\|g_{\zeta^{(r)}}^{\mathrm{tot}}\right\|_2^2\right)^{1/2}}{\left(N_{\mathrm{mask}}^{-1}\sum_r \left\|M_{\mathrm{eff}}\zeta^{(r)}\right\|_2^2\right)^{1/2}}, \tag{91}$$

when the denominator is positive. This ratio compares the root mean squared norms of the total residual and the complete linear term at the final representation, before normalization and multiplication by A. It is not a logit-error measure.

Numerical reconstruction errors. Table 1 reports the largest numerical discrepancy over all inputs, masks, seeds, values of q, and settings. The first row compares the recursive and matrix forms of the linear propagation. The second evaluates the bookkeeping identity that defines g_ζ^{tot}. In the two additive settings, this total residual is also the exact ReLU remainder from Theorem 18.

Table 1: **Numerical errors in linear propagation and reconstruction.** The values are maxima over the complete experiment.

Identity	maximum error
Linear propagation, $\widehat{\Delta}_L = M_{\mathrm{eff}}\zeta$	3.23×10^{-14}
Linear plus nonlinear reconstruction, $\Delta_L - M_{\mathrm{eff}}\zeta = g_\zeta^{\mathrm{tot}}$	2.48×10^{-14}

Predicted MSE reduction and sources of approximation error. Table 2 compares predicted and observed reductions after taking their medians within each experimental group. Pearson correlation measures association across these median pairs, while MAE averages their absolute differences. Correlations range from 0.9982 to 0.9997, with MAEs from 0.0254 to 0.0355. The table separately compares T_M^{-1} with the common scalar fitted on 128 masks. This coefficient comparison uses no evaluation masks.

To understand the remaining error in the predicted MSE reduction, let $\mathcal{E}_{\mathrm{aff}}$ be the relative MSE reduction for the linear model L_ζ^{aff}. Let $\mathcal{E}_{\mathrm{fr}}$ be the same reduction after replacing the random radius by $\bar{r}_M$. The column Δ_{nl} contains the median of $|\mathcal{E}_{\mathrm{obs}} - \mathcal{E}_{\mathrm{aff}}|$, while Δ_{rad} contains the median of $|\mathcal{E}_{\mathrm{aff}} - \mathcal{E}_{\mathrm{fr}}|$ within each group. After averaging over groups, the radius effect is about six to seven times larger than the nonlinear effect. For standard layerwise dropout, Δ_{nl} also includes the mask interactions in g_ζ^{tot}. Moreover, at the level of group medians, $\mathcal{E}_{\mathrm{fr}}$ differs from $\mathcal{E}_{\mathrm{pred}}$ by an average absolute error of 0.0016 to 0.0022 across settings. The reference-radius Monte Carlo calculation therefore agrees more closely with the exact formula than the calculation with a random radius does.

The input-level mean of $\mathcal{E}_{\mathrm{obs}} - \mathcal{E}_{\mathrm{pred}}$ is 0.0245, 0.0268, and 0.0355 for fixed branch outputs, standard layerwise dropout, and one earlier source, respectively. In these experiments, the reference-radius formula slightly underestimates the observed reduction.

Table 2: **Law relating the mean shift to dispersion after several ReLU layers.** Error in the predicted MSE reduction, scalar error, Δ_{nl}, and Δ_{rad} are reported as mean $\pm$ sample standard deviation over three seed summaries. Each seed summary averages its groups. Law MAE uses absolute differences of group medians. The next three errors use group medians of imagewise absolute differences. Correlations use all group medians. The last column is the median of the group medians of $\widehat{G}$.

Condition	$\mathrm{corr}(\mathcal{E}_{\mathrm{pred}}, \mathcal{E}_{\mathrm{obs}})$	law MAE	$\lvert\widehat{a} - T_M^{-1}\rvert$	Δ_{nl}	Δ_{rad}	$\widehat{G}$
Fixed branch outputs	0.9995	0.0254 ± 0.0019	0.00166 ± 0.00004	0.00306 ± 0.00016	0.02271 ± 0.00156	0.0468
Standard layerwise dropout	0.9997	0.0284 ± 0.0019	0.00140 ± 0.00011	0.00399 ± 0.00022	0.02255 ± 0.00148	0.1238
One earlier source	0.9982	0.0355 ± 0.0011	0.00832 ± 0.00102	0.00650 ± 0.00038	0.03755 ± 0.00019	0.0764

Standard layerwise dropout has a larger total residual than the setting with fixed branch outputs. Its median $\widehat{G}$ increases from 0.0468 to 0.1238, while its MAE of the predicted MSE reduction increases from 0.0254 to 0.0284. This experiment lies outside the exact additive assumptions, but it shows that the description in terms of mean shift and dispersion can remain useful when the interaction between masks stays moderate. Across all 105 groups, the Spearman correlation between $\widehat{G}$ and the median nonlinear effect $|\mathcal{E}_{\mathrm{obs}} - \mathcal{E}_{\mathrm{aff}}|$ is 0.883. Among the 90 groups with a positive effect, the Pearson correlation on logarithmic axes is 0.833.

Role of complete linear propagation. For fixed branch outputs, using only the last random source instead of the complete matrix M_{eff} increases MAE of the predicted MSE reduction from 0.0254 to 0.0582. For standard layerwise dropout, it increases the error from 0.0284 to 0.0612. Using the complete linear propagation reduces these errors by $56.4\,\%$ and $53.6\,\%$. For one earlier source, the effect depends on the number of later blocks. With four and eight later blocks, the complete matrix gives errors 0.0339 and 0.0266. Ignoring the later selected local maps gives 0.0792 and 0.1128. Using the complete matrix reduces these errors by $57.1\,\%$ and $76.4\,\%$. With no later block, the two matrices are equal.

With one or two later blocks, the incomplete scalar approximation yields smaller observed errors in the predicted MSE reduction. Its errors are 0.0033 and 0.0342, compared with 0.0411 and 0.0406 for the complete matrix. This reversal concerns the final scalar summary and is consistent with partial cancellation between approximation errors. The complete vector propagation still uses all selected local maps. Its recursive and matrix forms agree up to a maximum numerical discrepancy below 3.23×10^{-14} over the complete experiment.

Finite-dimensional bounds. Proposition 19 bounds the union event that at least one ReLU unit changes state. Each network contains hundreds of such units, and this event occurs for almost every tested setting with at least one later ReLU block. Many individual changes have almost no effect on the final representation. The ratio $\widehat{G}$ measures the remaining perturbation there, whereas the union event records any state change. Writing $\widehat{\eta}_M$ for the 160-mask plug-in estimate, the condition $\widehat{\eta}_M \le 1/4$ holds for $6\,418$ of the $53\,760$ records. The corresponding plug-in inequality, evaluated on the same 160 masks, holds for every one of these records. The other $47\,342$ records do not satisfy the plug-in condition and are excluded from this evaluation of the inequality. This is an in-sample numerical evaluation in which population mean squared errors are replaced by averages over the same masks. It does not validate the corresponding population bound. Theorem 20 is a stability inequality, not a confidence-coverage statement. Among the $6\,418$ eligible records, the median right side is $12\widehat{\eta}_M = 2.36$, so the numerical bound is conservative. Theorem 20 provides the finite-dimensional guarantee in (87), while $\widehat{G}$ reports the measured size of the residual relative to the linear perturbation before normalization.

Overall, the experiment evaluates the complete linear propagation, quantifies the residual after several ReLU layers, and finds that the law relating the mean shift to dispersion closely matches the observed group summaries. Radius fluctuations cause most of the residual discrepancy, while standard layerwise dropout shows that moderate interactions between masks can still leave the same description useful beyond the exact additive assumptions.

E EXPERIMENTAL DETAILS

This appendix explains how the theoretical quantities are estimated and how each experimental setting is built. We begin with the common estimators, then describe the MNIST, language model, and FaceNet protocols in the order in which they appear in the paper. Practical reproduction instructions are in `code_dropout.zip/README.md`.

E.1 COMMON ESTIMATORS

All logit-space quantities are computed after removing the common logit shift. For m outputs, we apply $C_m = I_m - m^{-1}\mathbf{1}_m\mathbf{1}_m^\top$. For language models, m is the vocabulary size. This centering is necessary because softmax is invariant to a common shift and the theory concerns only identifiable logit directions.

One local experimental point. For a fixed input, we run the trained model once up to the selected residual branch. We store $h = x + u$, the dropout-affected term u, the output map A, the fixed term c, and the keep probability q. These quantities determine T, ℓ_T, and the predicted relative MSE reduction. Only after this calculation do we draw new masks in the selected branch and measure the resulting outputs. The preceding activations and the model weights remain fixed. A condition corresponds to one model, one keep probability, and one specified intervention. Its reported point is obtained by applying the same summary to the predicted and observed values over inputs. This order is important. The random masks assess the prediction computed from the stored representation.

For input i and masks $j \in [\![1, N_{\mathrm{mask}}]\!]$, let ℓ_{ij} be the centered random logits, ℓ_{i0} the centered original logits, and ℓ_{iT} the centered theory target. When the denominator is positive, the observed relative MSE reduction is

$$\widehat{\mathcal{E}}_i = \frac{\sum_{j=1}^{N_{\mathrm{mask}}} \|\ell_{ij} - \ell_{i0}\|_2^2 - \sum_{j=1}^{N_{\mathrm{mask}}} \|\ell_{ij} - \ell_{iT}\|_2^2}{\sum_{j=1}^{N_{\mathrm{mask}}} \|\ell_{ij} - \ell_{i0}\|_2^2}. \tag{92}$$

The denominator is the total squared change produced by dropout relative to the output without dropout. The numerator is the decrease in this error when ℓ_{i0} is replaced by the analytic target ℓ_{iT}. Therefore, $\widehat{\mathcal{E}}_i = 1$ means that the target removes the whole change, $\widehat{\mathcal{E}}_i = 0$ means that it gives no reduction, and a negative value means that it increases the squared error.

For the same input, we compute T_i. When $\ell_{i0} - c \neq 0$, we also compute $\mathcal{V}_i$. If $\mathcal{V}_i > 0$, we then compute Λ_i and $\mathcal{E}_{i,\mathrm{pred}} = \Lambda_i/(1+\Lambda_i)$. Group points in Figure 1 use matching medians for predicted and observed reductions. For MNIST and FaceNet, we take medians over inputs. For language models, we first take medians over token positions within each passage, then over passage summaries. The pooled native summary uses twenty passages, ten from validation and ten from test. Each keep-probability group uses ten validation passages. Appendix F.5 also evaluates prediction error over individual test tokens.

Scalar-fit diagnostics. The MNIST projection diagnostic asks how much of each random output lies outside the affine line $c + \mathbb{R}(\ell_0 - c)$. When $\ell_0 - c \neq 0$, it fits a separate scalar for each mask. The "best-scalar residual" divides the norm of the orthogonal remainder by $\|\ell_\zeta - c\|_2$, when this denominator is nonzero. The numerator term $APD\zeta/r_\zeta$ in Proposition 1 may also contain a component along the direction $\ell_0 - c$.

The local language model diagnostic instead asks how well one scalar approximates all masks of a token. For input i, define the empirical mean squared error and its minimizing coefficient by

$$\widehat{\mathcal{Q}}_i(a) := \frac{1}{N_{\mathrm{mask}}} \sum_{j=1}^{N_{\mathrm{mask}}} \|\ell_{ij} - \{c + a(\ell_{i0} - c)\}\|_2^2, \qquad \hat{a}_i := \arg\min_{a \in \mathbb{R}} \widehat{\mathcal{Q}}_i(a). \tag{93}$$

The fit is unique when $\ell_{i0} - c \neq 0$. To express its error relative to the size of the analytic output component that depends on the features, we report the normalized fitted root mean squared error

$$\frac{\sqrt{\widehat{\mathcal{Q}}_i(\hat{a}_i)}}{\|(\ell_{i0} - c)/T_i\|_2}.$$

FaceNet uses the same input-wise fit and reports the square of this normalized RMSE. These fits are diagnostics, not inputs to the analytic prediction. The MNIST common-scalar comparison uses the same minimization on 128 fitting masks and evaluates the resulting coefficient on 512 independent masks (160 for ReLU). The minimizer on the evaluation masks is recorded separately. It measures in-sample fit and is not the independently evaluated coefficient. The projection diagnostic instead takes the lower median over masks within each image, then the usual median over images. For MSE-based condition summaries, the usual median averages the two central values when necessary.

To measure relative fluctuations of the normalization radius, define

$$\mathrm{CV}(r_\zeta) \coloneqq \frac{\sqrt{\mathrm{Var}(r_\zeta)}}{\mathbb{E} r_\zeta}.$$

We estimate both moments by averaging over masks.

All MNIST squared errors are computed directly from residual vectors in double precision. Undefined ratios are recorded as missing with their denominators, rather than assigned zero. None of the 492 000 local condition/image pairs has zero signal, an unresolved baseline error, or an active normalization floor. Nonpositive projection coefficients remain in the unconstrained fits. Paired image bootstraps use 2 000 resamples for the observed condition median minus the predicted condition median, conditional on each checkpoint. They do not quantify variation across training seeds.

For the intervals comparing predicted and observed reductions in pretrained language models, each text block is one resampling unit. We draw the ten block summaries with replacement 3 000 times and recompute the paired difference $\mathrm{median}(\mathcal{E}_{\mathrm{obs}}) - \mathrm{median}(\mathcal{E}_{\mathrm{pred}})$. This preserves dependence among tokens from the same context. Simple median intervals use 2 000 resamples. The paired Monte Carlo comparisons, scalar stability comparisons, and intervals for dropout throughout the network use 5 000 resamples at the passage level.

E.2 MNIST ARCHITECTURE AND TRAINING DETAILS

The 27 local models use normalized MNIST images and a common 30 000-image training subset. A permutation of the official training split with seed 2026 assigns its first 30 000 images to training and its last 10 000 to validation. The three deep ReLU models use all 50 000 non-validation images, ordered with seed 20260803. Test images enter neither training nor checkpoint selection. We retain every architecture, width, and seed, independently of its agreement with the theoretical prediction.

AdamW uses weight decay 10^{-4}, batch size 256, and a cosine learning-rate schedule. Local models have a maximum of ten epochs, at initial learning rate 2×10^{-3} for residual MLPs and 1.5×10^{-3} for Mini-ViT and PreAct ResNet. Their seeds are 11, 17, and 23, with training keep probability $q = 0.65$. Deep ReLU models use seven epochs, learning rate 2×10^{-3}, gradient clipping at norm 5, and seeds 0, 1, and 2. Early stopping monitors validation cross-entropy with patience three and minimum improvement 10^{-4}. The retained checkpoint has the smallest validation cross-entropy, with the earliest epoch breaking exact ties. All runs reach their maximum epoch budget.

Training uses single precision on a CPU. Diagnostics from stored representations use double precision. Main probes use the first 2 000 official test images, transformations use the first 800, and ReLU probes use the first 512. Table 3 reports classification performance on the complete test set, separate from the mechanism diagnostics. The MNIST test benchmark had already been examined during this project, so it is not a new external validation set.

Table 3: Classification performance at the validation-selected checkpoints. Entries are ranges over all models in each family. Validation and test each contain 10 000 images. Validation is used for selection.

Family	models	validation accuracy (%)	test accuracy (%)
Residual MLP	9	96.96 to 97.85	97.13 to 98.04
Mini-ViT	9	94.31 to 96.48	94.67 to 96.79
PreAct ResNet	9	98.73 to 99.38	98.82 to 99.44
Deep ReLU	3	97.84 to 98.00	97.98 to 98.20

Table 4: Hand-built MNIST architectures. The last listed branch is the one resampled in the local experiments.

Model	Main components	terminal dimensions	final geometry
Residual MLP	linear stem, two-layer GELU branch, residual addition	64, 128, 256	RMSNorm
Mini-ViT	7×7 patches, two blocks, four attention heads, GELU MLP branch	32, 64, 96	LayerNorm
PreAct ResNet	four pre-activation blocks, last convolutional branch, global average pooling	32, 64, 96	ℓ_2 norm

The ResNet normalizes a representation v as $v/\max(\|v\|_2, 10^{-8})$. Above the norm floor, this is the $\varepsilon = 0$ convention discussed in Appendix C. The mini-ViT keeps its class-dependent output bias and LayerNorm shift. The analytic target is therefore ℓ_T in (8). A separate stress test multiplies the fixed term so that its norm is between zero and twice the median variable-logit norm. The fixed target remains valid, while dividing all logits by T becomes increasingly inaccurate.

The native grid evaluates $q \in \{0.50, 0.65, 0.80, 0.90\}$ with 512 evaluation masks per image, for 108 conditions. No condition exceeds the specified 0.01 threshold for the change in median relative MSE reduction between the 256-mask and 512-mask prefixes. This is a sensitivity analysis, not a guarantee for the infinite-mask mean. On the widest model of each family and all three seeds, we add the following comparisons at $q = 0.65$: Gaussian coordinate noise with variance ρ, channel-wide DropPath, dropout after normalization, removal of terminal normalization, and a concentrated output matrix. The main probabilistic assumptions apply to independent Bernoulli and Gaussian noise before normalization. The other settings probe the geometric boundaries.

For the study of replication preserving the original function, we use the widest model of each family, 800 test images, 128 fitting masks, and 512 evaluation masks. Lift factors are $k \in \{1, 2, 4, 8\}$. Independent masks use $q \in \{0.50, 0.65, 0.80\}$, giving 108 conditions. Shared masks use $q = 0.65$, giving 36 conditions, and are coupled across copy factors to verify invariance. Both settings use the same original lifted representation. For RMSNorm and LayerNorm, the implementation stabilizer is divided by k, keeping the additive term in the squared Euclidean radius unchanged. The rotation test of Theorem 8 uses only the residual MLP with RMSNorm and the ResNet with Euclidean normalization. It uses the identity, three Haar draws, and one output-aligned construction for each model and seed. These models have $P = I_p$, as required by the theorem.

A separate LayerNorm experiment uses the Mini-ViT at dimension $p = 96$ for all three seeds. We use 800 test images, 512 evaluation masks, and $q = 0.65$. Write $e = p^{-1/2}\mathbf{1}_p$ and let U_c be an orthonormal basis of $e^{\perp}$. Every valid random rotation has the form $Q = ee^{\top} + U_c R U_c^{\top}$, where R is Haar distributed on $O(p-1)$. We test the identity, five independent draws of R for each seed, and a centered SVD construction that makes fewer centered coordinates directly affect the output numerator. An unrestricted Haar rotation is included to show what fails when $Q\mathbf{1}_p \neq \mathbf{1}_p$. All rotations use the inverse change in the output matrix and require no retraining.

E.3 Pretrained language model protocol

What is resampled in a language model. The local experiment does not activate dropout throughout the transformer. For each token, we first run the pretrained model once and record the fixed residual input x and the feed-forward output u at the selected branch. We then apply a new inverted-dropout mask to u, form the corresponding residual sum, and pass it through the following normalization and vocabulary projection. Repeating only this last operation gives several random logit vectors for the same token. Earlier blocks are not recomputed with new masks. This construction isolates the setting of the local theorem. Dropout throughout the complete model is studied separately as a boundary test.

Every checkpoint enters every language model experiment. The experiments include the native validation and test analyses, the five-value keep-probability sweep, and the independent MC-2048 comparison. They also include the branch-position and matched-Gaussian comparisons, removal of

Table 5: Pretrained language models. Parameter counts include the output matrix. For T5 and BART, the table lists decoder blocks.

Model	parameters	blocks	dimension	vocabulary
GPT-2	124.4M	12	768	50 257
GPT-2 Medium	354.8M	24	1 024	50 257
GPT-2 Large	774.0M	36	1 280	50 257
GPT-2 XL	1.56B	48	1 600	50 257
OPT-125M	125.2M	12	768	50 272
OPT-350M	331.2M	24	1 024	50 272
OPT-1.3B	1.32B	24	2 048	50 272
T5 Small	60.5M	6	512	32 128
T5 Base	222.9M	12	768	32 128
BART Base	139.4M	6	768	50 265

the fixed residual input, dropout throughout the network, and a comparison of two scalar estimates on new masks.

We use the public pretrained checkpoints and WikiText-2 raw validation and test splits. GPT-2 and OPT are evaluated autoregressively. For T5 and BART, each text window is split into a 64-token encoder prefix and a 64-token decoder continuation. Teacher forcing supplies earlier target tokens. We discard the first eight output positions. The same ten passages are used for every model and each split.

For each checkpoint, we select the last feed-forward residual branch followed by a final LayerNorm or RMSNorm that feeds the vocabulary projection. This is the final model normalization for GPT-2 and the pre-normalization OPT models. OPT-350M and BART use the last block normalization. T5 uses the final decoder RMSNorm. The representation immediately before this normalization is written as $h = x + u$. We verify that $x + u$ reconstructs the captured representation. We also verify that the normalization and vocabulary projection reconstruct the model logits. The residual reconstruction error is zero in the stored precision. The largest logit difference is 0.25 on one GPT-2 Medium output coordinate. This is an FP16 rounding difference among 50 257 outputs. Normalization gains, shifts, output projections, and the BART output bias are included through A and c.

The native evaluation uses 120 positions on validation and 120 on test, with 12 positions from each passage and 32 masks. All checkpoints have native keep probability $q = 0.9$. The keep-probability sweep uses all ten checkpoints. For each $q \in \{0.50, 0.70, 0.80, 0.90, 0.95\}$, it uses 80 positions, eight per passage, and 32 paired masks. The resampling unit is a full passage. Intervals for individual group medians use 2 000 resamples of the ten passages. Intervals for the difference between observed and predicted median reductions use 3 000 resamples, as specified in Appendix E.1.

The probability experiment uses 40 positions, four per passage. Two disjoint sets of 1 024 masks form the MC-2048 target. We compare the deterministic output, the analytic one-pass output, sampled averages with 8, 16, 32, or 64 masks, and a scalar fitted on the same number of masks. We report the difference between the two MC-1024 halves because the finite MC-2048 target still has sampling error. A final experiment masks the first, middle, or last feed-forward branch in every checkpoint. It uses four positions from each passage and 16 masks. Since later blocks can change the geometry, this experiment fits a scalar to the complete centered final logits, as detailed in Appendix F.4.

The matched Gaussian and fixed-residual experiments use 80 validation positions, eight from each passage, and 32 masks. For the experiment with dropout throughout the network, we place the complete model in training mode and keep its pretrained weights fixed. This activates the native dropout operations for 16 passes at two positions from each passage. In the scalar comparison, one scalar is fitted on 24 masks. We compare it with the explicit $1/T$ on 48 new masks, using 80 positions with eight from each passage.

E.4 Pretrained angular face-embedding protocol

We use the public InceptionResnetV1 FaceNet implementation with VGGFace2 weights (Schroff et al., 2015; Cao et al., 2018). The network is fixed and evaluated on all 6 000 pairs in the canonical

LFW verification file (Huang et al., 2007): 3 000 same-identity and 3 000 different-identity pairs, organized in the ten official folds. These pairs reference 7 701 unique funneled images. The first image receives a random mask. The second image is its fixed gallery partner. This matches a common use case in which stored gallery embeddings are compared with new images. The theory uses 4 564 distinct images and no identity labels.

Each referenced image crosses the convolutional backbone once on a T4 GPU. We resize the standard funneled image deterministically to 160×160 and cache the 1 792-dimensional pooled feature g immediately before the model's native dropout. In evaluation mode, the following linear layer and BatchNorm form one affine map,

$$h = W_{\text{eff}} g + b_{\text{eff}}, \qquad e = \frac{h}{\max(\|h\|_2, 10^{-12})}.$$

The reconstructed normalized embedding agrees with the mixed-precision full model with a maximum absolute error of 1.39×10^{-4}. Under an inverted dropout mask, the random tail is exactly

$$h_\zeta = h + M_g \zeta, \qquad M_g = W_{\text{eff}} \operatorname{diag}(g),$$

This additive perturbation has the rectangular form of Proposition 11, with $d = 1792$ independent sources and normalized dimension $p = 512$. Normalization uses the same norm floor for h_ζ. Above that floor, it follows the zero-stabilizer convention discussed in Appendix C.

The native dropout probability is 0.6, hence the native keep probability is $q = 0.4$. The primary sweep uses $q \in \{0.40, 0.55, 0.70, 0.85, 0.95\}$ and 48 masks per test image. A Gaussian source with the same covariance is evaluated at $q = 0.4$. The Bernoulli sweep reuses the same underlying uniform draws across keep probabilities. The native run also fits scalars on masks 1-16 and 1-24, evaluates the latter on the different masks 25-48, and records the empirical optimum over all 48 masks.

The dependence ablation uses one shared Bernoulli variable per randomly permuted group of 8, 32, 128, or 1 792 source coordinates. These settings leave 224, 56, 14, or 1 independent groups, respectively. They use 32 masks on 1 024 randomly chosen test images and a fixed 256-image gallery. The original linear map and every original embedding remain fixed. A comparison uses 1, 32, 128, or 512 visible gallery scores.

For the primary experiment comparing predicted and observed MSE reductions, the output map A contains 512 unit gallery embeddings sampled once from the second images in the official pairs. One output is therefore the complete vector of cosine similarities between a test image and this fixed diagnostic gallery. The verification experiment instead uses all 6 000 paired scores. For each official fold, a cosine threshold is selected on the other nine folds. We report both a common threshold learned from deterministic scores and a method-specific threshold. Labels enter only these verification metrics. For the primary intervals comparing predicted and observed MSE reductions, we resample the 4 564 test images 2 000 times, keeping the diagnostic gallery fixed. Timing for the final layers uses the same 256 stored images and does not include the common feature extractor, gallery products, or CSV creation.

The independent experiment selects 512 official pairs while keeping the two identity classes balanced inside each fold. Six folds contribute 26 pairs per class and four folds contribute 25. Two disjoint sets of 1 024 masks give independent means, and their average is the MC-2048 target. The MC-16 and MC-48 baselines use ten independent repeats. MC-128 uses six. None of these masks enters the target. We estimate uncertainty by 2 000 paired bootstrap draws that keep all scores from the same first image together. The reported MC-2048 noise estimate is half the mean absolute difference between the two MC-1024 references under the Gaussian sampling approximation.

F Additional results

We first compare the law relating the mean shift to dispersion and scalar coefficients across all settings, then turn to the questions specific to MNIST, language models, and FaceNet. Table 6 provides a guide to this progression.

Table 6: Guide to the additional experiments.

Setting	Additional tests	Location
MNIST models	changes of coordinates, independent and shared masks, rotations compatible with LayerNorm, one-pass probabilities, and several ReLU layers	F.3, D.2, F.7
FaceNet	five keep probabilities, Gaussian noise, independent mask groups, normalization, visible gallery scores, and an independent MC-2048 target	F.6
10 language models	five keep probabilities, token-level prediction, independent MC-2048 targets, branch position, Gaussian noise, removal of the fixed residual, and dropout throughout the network	F.4, F.5, F.7

Table 7: Fits of predicted and observed MSE reductions across groups using matching median aggregation for predicted and observed relative MSE reductions. The regression is $\mathcal{E}_{\mathrm{obs}} = a + b\mathcal{E}_{\mathrm{pred}}$.

Subset	groups	correlation	MAE	b	a	max error
MNIST native and independent copies	216	0.99946	0.01806	1.0268	0.00985	0.03845
All 10 language models, native q	10	0.99123	0.00022	1.1643	−0.00017	0.00055
All 10 language models, five q values	50	0.98336	0.00063	0.9089	0.00007	0.00834
All 10 language models, Gaussian noise	10	0.98359	0.00020	0.9658	−0.00003	0.00045
All 10 language models, no fixed residual	10	0.92430	0.00209	0.8604	0.00332	0.00950
Pretrained FaceNet embedding	6	0.99998	0.00040	0.9425	0.00000	0.00081

F.1 Predicted and observed MSE reductions

Each row summarizes a set of pairs $(\mathcal{E}_{\mathrm{pred}}, \mathcal{E}_{\mathrm{obs}})$, one pair for each experimental condition in that row. The fitted line is only a description of this set of points. It is not used to define T, ℓ_T, or any prediction reported in the paper.

The identity line corresponds to $a = 0$ and $b = 1$. Correlation describes the association between predicted and observed values across conditions. MAE and maximum error test absolute agreement, while a and b expose additive and multiplicative deviations from the identity line. A high correlation alone is therefore not treated as numerical agreement.

The MNIST row includes all 108 native conditions and all 108 independent-copy conditions. Rotations and boundary controls are examined separately below. Shared copies use the formula accounting for mask covariance and are not pooled with the independent-copy comparison. The native row uses one median pair for each checkpoint in Table 5. The next three rows use every checkpoint. They cover five Bernoulli keep probabilities, matched Gaussian noise, and a setting where the fixed residual input is removed. The last setting makes the random branch much stronger, so its larger error marks the limit of the local approximation.

F.2 Empirical scalar comparisons

Figure 1(c) compares $a_{\mathrm{theory}} = 1/T$ with scalar coefficients computed from random outputs. For each MNIST image in the independent-copy study, a_{fit} minimizes mean squared error over 128 fitting masks. It is the mean of their projection coefficients, without a positivity constraint. We take matching image medians of a_{fit} and $1/T$. For each language model token, a_{fit} is the least-squares coefficient. We then take one median coefficient pair for each model and keep probability. For FaceNet, both coefficients pool the exact quadratic statistics over 4 564 test images and 512 fixed gallery scores. Labels are absent from every fit.

The complete language model run also gives a direct test on new masks, using the empirical risk in (93). For each token i, we fit $\hat{a}_i$ on 24 masks and evaluate both $\hat{a}_i$ and $1/T_i$ on 48 new masks. Thus the fitted scalar cannot reuse the randomness that measures its final error. This held-out comparison separates estimation from assessment. A scalar can fit the first 24 masks partly because of finite-sample variation. Its error on 48 different masks measures whether it remains suitable for new dropout draws. The analytic coefficient $1/T_i$ uses no sampled mask and is assessed by exactly the same rule.

Table 8: Agreement between the explicit scale $1/T$ and empirical scalar coefficients. Errors are absolute differences between group-level coefficients.

Setting	groups	correlation	MAE	maximum error
Replication with independent masks	108	0.99999	1.98×10^{-4}	8.11×10^{-4}
All 10 language models, native q	10	0.98573	1.97×10^{-4}	4.57×10^{-4}
All 10 language models, five q values	50	0.99388	5.38×10^{-4}	4.30×10^{-3}
Pretrained FaceNet scores	5	> 0.99999	4.52×10^{-4}	1.12×10^{-3}

Table 9: Scalar comparison on new masks for every language model. Excess risk is the pooled increase above the empirical minimum on the 48 evaluation masks, divided by the pooled MSE of the output without scalar correction. Entries are percentages.

Model	explicit excess	24-mask-fit excess	explicit wins
GPT-2	0.5324 %	1.1813 %	71.3 %
GPT-2 Medium	0.2248 %	0.5945 %	68.8 %
GPT-2 Large	0.0459 %	0.2058 %	71.3 %
GPT-2 XL	0.0617 %	0.1738 %	80.0 %
OPT-125M	0.0072 %	0.0188 %	80.0 %
OPT-350M	0.0203 %	0.0565 %	68.8 %
OPT-1.3B	0.0109 %	0.0246 %	75.0 %
T5 Small	0.2326 %	0.8091 %	71.3 %
T5 Base	0.1504 %	0.3925 %	68.8 %
BART Base	0.0364 %	0.1593 %	70.0 %

The explicit coefficient has lower pooled excess risk for every checkpoint. It also wins on $68.8\,\%$ to $80.0\,\%$ of tokens. For each model, a paired bootstrap over the ten passages gives a positive 95 % interval for the fitted excess minus the explicit excess.

The FaceNet run gives the same type of test at its native keep probability $q = 0.4$. We fit $\hat{a}_{1:24}$ on masks 1-24, then evaluate both $\hat{a}_{1:24}$ and the label-free $1/T$ only on masks 25-48. Relative to the empirical optimum computed on these last 24 masks, the excess risks are $0.0918\,\%$ for $\hat{a}_{1:24}$ and $0.0501\,\%$ for $1/T$. Thus the explicit coefficient reduces the excess by $45.4\,\%$ and has smaller held-out risk on $62.1\,\%$ of the 4 564 test images. Resampling the images gives a 95 % interval $[42.9\,\%, 47.9\,\%]$ for the reduction in excess. These excess risks are divided by the pooled risk of the empirical optimum on masks 25-48, rather than by the uncorrected baseline risk used for the language models.

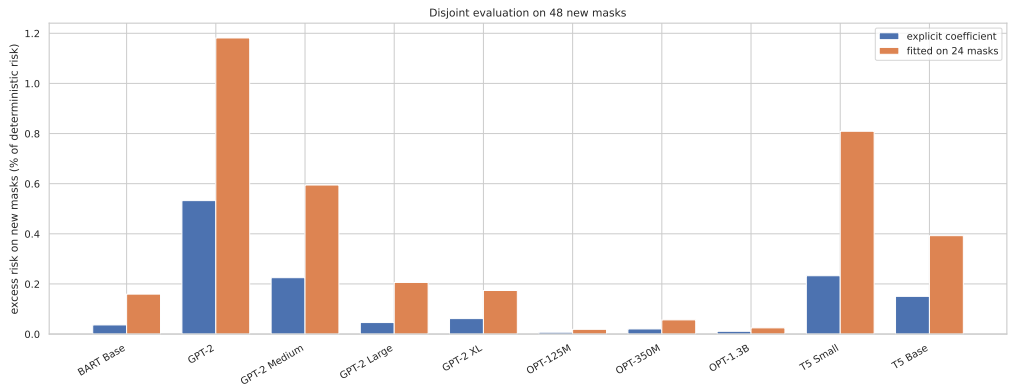


Figure 2: **The explicit coefficient is stable on new masks.** The fitted scalar uses 24 masks. Both scalars are then evaluated on 48 different masks. Excess risk uses the uncorrected baseline MSE as its denominator, as in Table 9. Lower values are better.

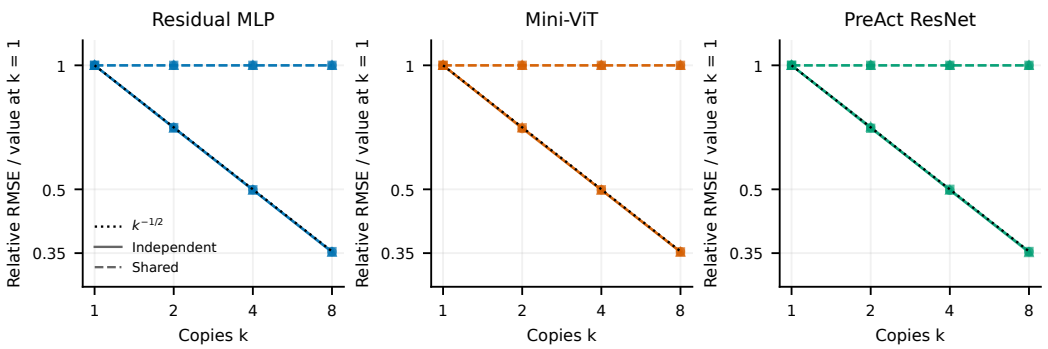


Figure 3: **Independent copies reduce the error relative to the analytic target.** For each image, RMSE is divided by $\|(\ell_0 - c)/T\|_2$. Each curve is the median over 800 images for one training seed, divided by its value at $k = 1$. All nine widest models use $q = 0.65$ and 512 evaluation masks. Solid curves use independent masks and dashed curves use shared masks. The dotted reference is $k^{-1/2}$. Markers distinguish the three seeds.

F.3 Additional experiments on MNIST models

The MNIST experiments below study changes of coordinates that preserve the original predictions. They complement the independent-copy experiment in Figure 3. Table 18 later reports the one-pass probability approximation. Appendix D.2 gives a separate MNIST study with eight ReLU blocks, 53 760 records, and 105 experimental groups.

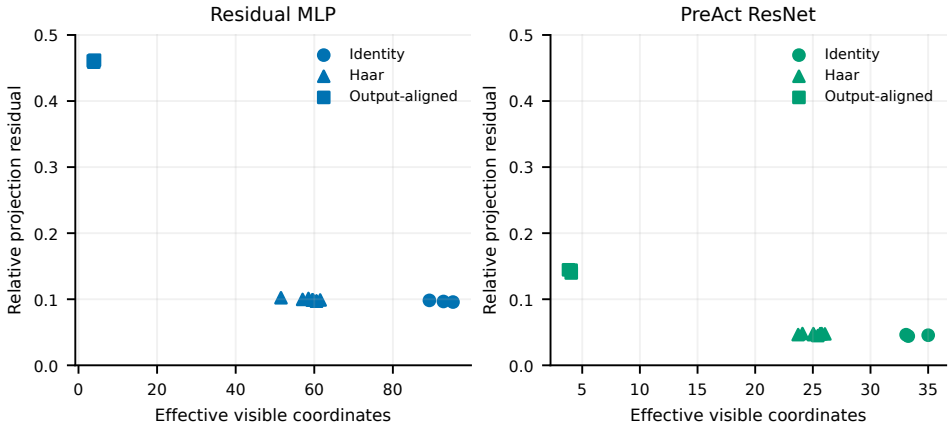


Figure 4: **Coordinate dropout is basis dependent.** Each rotation is paired with the inverse rotation of the output matrix, so the original classifier is preserved. The figure contains the two models with $P = I_p$, as required by Theorem 8. Each point is an image median for one seed and one rotation. All three seeds and all three Haar draws per seed are shown. In these models, output-aligned coordinates reduce the effective number of visible coordinates and increase the projection residual for each mask.

Prediction-preserving rotations. Across every lift and displayed rotation, the largest absolute difference between the original logits and the reconstructed logits is below 5.0×10^{-14} in double precision. This is below the stated numerical tolerance 10^{-4}. The independent lift exponents at $q = 0.65$, averaged over the nine architecture and seed pairs, are shown below.

Table 10: Slopes on logarithmic axes over $k \in \{1, 2, 4, 8\}$ at $q = 0.65$. Mean and sample standard deviation use all nine architecture/seed pairs.

Mask dependence	metric	mean slope	standard deviation
independent	relative RMSE against the analytic target	-0.501377	0.001267
independent	visible ratio $\mathcal{V}$	-1.000000	$< 10^{-12}$
shared	relative RMSE against the analytic target	$< 10^{-12}$	$< 10^{-12}$
shared	ratio accounting for mask covariance $\mathcal{V}_{\Sigma,k}$	$< 10^{-12}$	$< 10^{-12}$

For positive metrics, these log-log regressions use the identity $\log y = \log C + b \log k$ when $y = Ck^b$. The slope therefore estimates the power-law exponent b, while the intercept estimates $\log C$. Exponentiating the intercept gives the multiplicative constant C. For independent masks, the relative RMSE against the analytic target has a slope close to $-1/2$. This is the error controlled by Theorem 7, not the projection residual for each mask used in the rotation experiment. Four copy factors provide evidence at the tested copy factors, not proof of an asymptotic rate. The exact identity $\mathcal{V}_k = \mathcal{V}_1/k$ gives slope -1 when $\mathcal{V}_1 > 0$. Shared masks preserve both the residual and the ratio accounting for mask covariance $\mathcal{V}_{\Sigma,k}$ in (53). The independent-coordinate expression $\mathcal{V}_k$ is not a shared-mask variance and retains its algebraic $1/k$ scaling. The measured shared-mask slopes are close to zero.

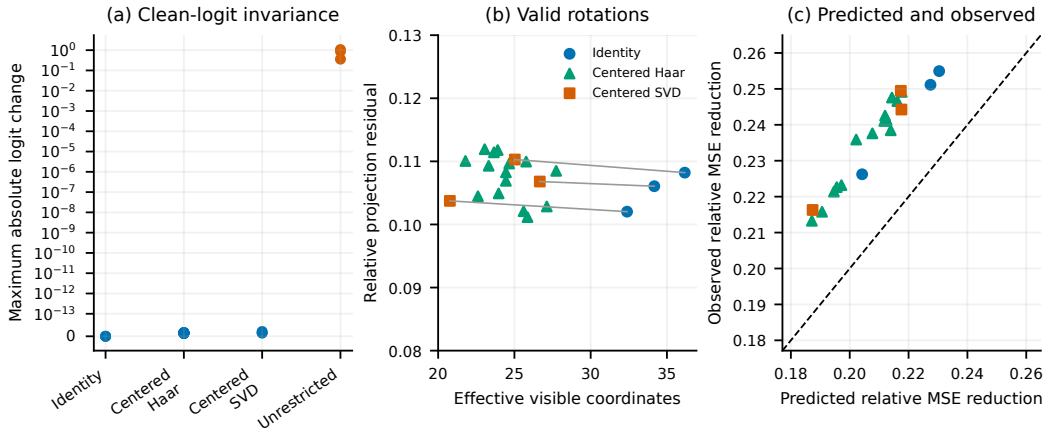


Figure 5: **LayerNorm allows rotations inside the centered subspace.** (a) The identity, centered Haar rotations, and the centered SVD construction satisfy $Q\mathbf{1}_p = \mathbf{1}_p$ and preserve the original logits. An unrestricted Haar rotation changes the classifier. (b) Valid rotations change how many coordinates are visible to the output and the projection residual for each mask. Thin lines join the identity and centered SVD result for the same seed. (c) Predicted and observed reductions for all 21 valid groups. Panels (b) and (c) report medians over images. Panel (a) uses a symmetric logarithmic scale with a linear region below 10^{-13}, so exact zero remains visible.

Rotations compatible with LayerNorm. Proposition 17 gives the reason for the condition $Q\mathbf{1}_p = \mathbf{1}_p$. It makes the rotation commute with LayerNorm centering. Across all three Mini-ViT seeds, the largest logit change under a valid rotation is below 2.14×10^{-14}. The median, over the three seeds, of the maximum logit change under an unrestricted Haar rotation is 0.876. Thus a generic rotation in $O(p)$ changes the LayerNorm classifier.

For the 21 valid groups, predicted and observed relative MSE reductions have correlation 0.969 and MAE 0.0281. The centered rotations produce a modest change in this trained Mini-ViT rather than a different regime. Their main role is to verify the LayerNorm symmetry and to show that the same law relating the mean shift to dispersion remains a useful approximation after the allowed changes of coordinates. Mean d_{eff} is 34.24 for the identity, 24.53 for centered Haar rotations, and 24.15 for the centered SVD construction. These are means of image medians over all seeds and rotation draws. The corresponding projection residuals are 0.1054, 0.1076, and 0.1069. The small, nonmonotone changes do not establish that fewer effective coordinates always increase the residual.

F.4 Additional experiments on ten language models

The larger language model study separates stability from dominance. Within a model family, a larger feature dimension often stabilizes both the radius and the residual output perturbation. This does not make the analytic scalar dominant at the locations studied here because the systematic scalar shift also becomes small.

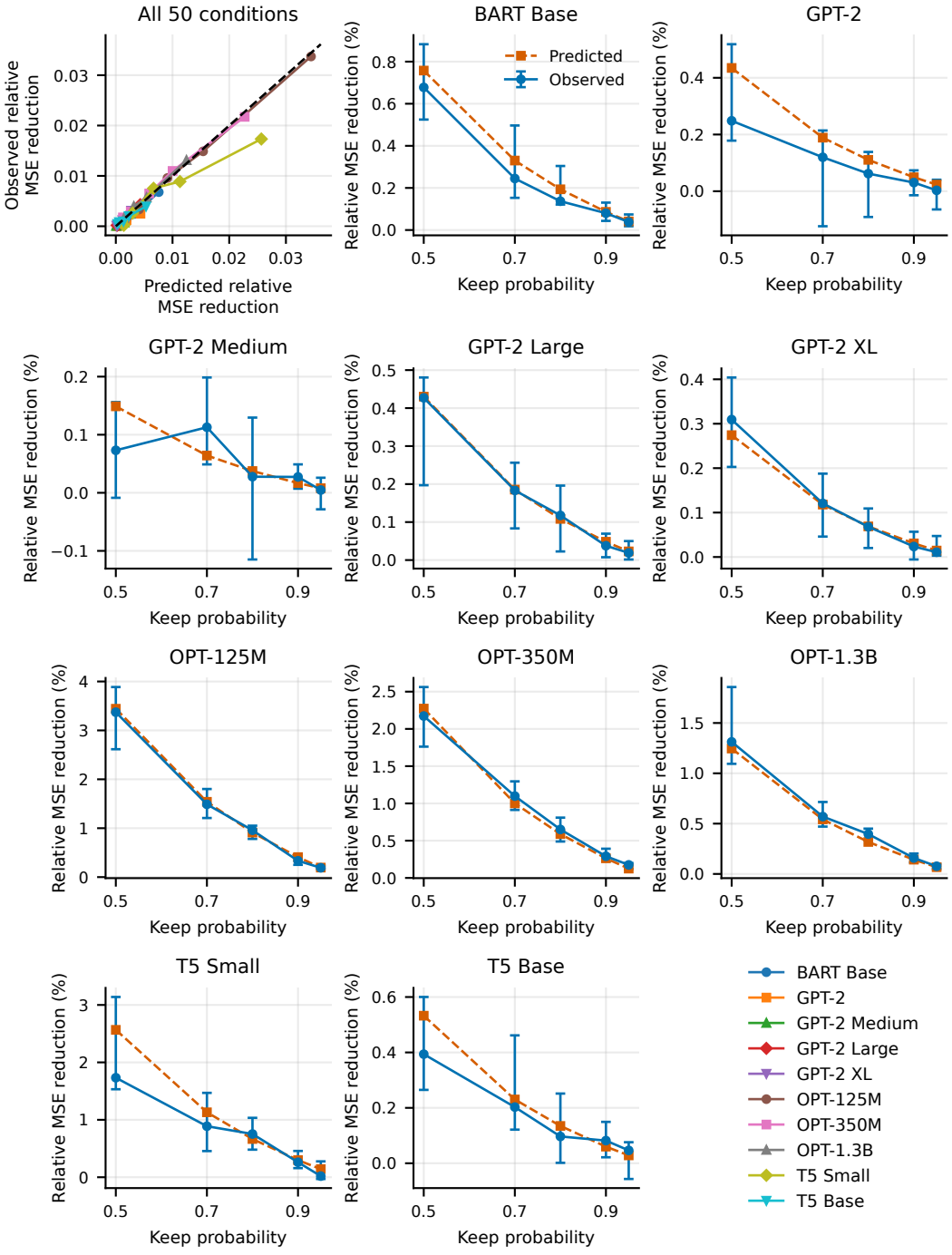


Figure 6: **The law relating the mean shift to dispersion across all ten language models.** Each panel uses five keep probabilities. Points are medians of passage summaries and bars are 95 % intervals from resampling the ten passages. The first panel collects all fifty model and keep-probability groups.

Across the 50 keep-probability groups in Figure 6, the predicted and observed relative MSE reductions have correlation 0.98336 and MAE 6.31×10^{-4}. At the native keep probability, $\sqrt{\mathcal{V}}$ and the normalized fitted RMSE defined in Appendix E.1 have correlation 0.99943 across the ten models, with MAE 5.58×10^{-4}. Thus the law relating the mean shift to dispersion predicts the relative MSE reduction, while $\mathcal{V}$ predicts the residual scale left by a scalar.

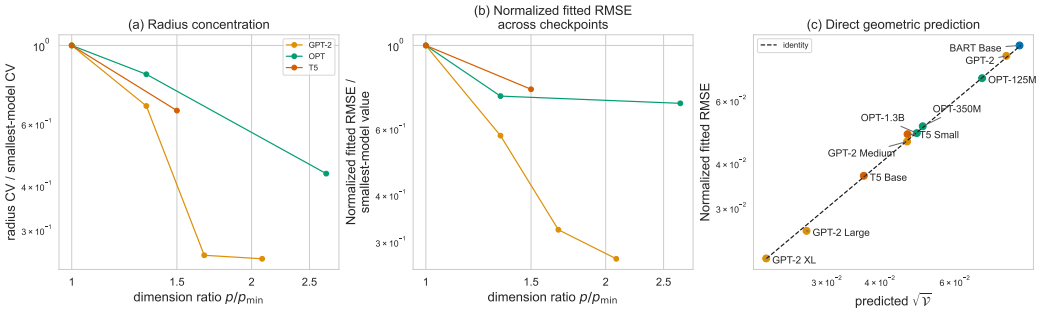


Figure 7: **Feature dimension and local geometry.** Each point is one pretrained checkpoint at the native keep probability. Lines join models from the same family and summarize the available checkpoints. Panel (b) divides the normalized fitted RMSE by its value for the smallest model in each family. Panel (c) uses the normalized fitted RMSE directly.

The within-family lines are descriptive because the checkpoints were trained separately. The replication preserving the original function in Appendix F.3 isolates the effect of dimension while holding the original network function fixed.

Table 11: Native results from 20 passage blocks per model, ten from validation and ten from test. Entries are medians of medians within passages. Every entry after the dimension is a percentage. Fitted RMSE is normalized by $\|(\ell_{i0} - c)/T_i\|_2$, as in Appendix E.1. The slightly negative observed $\mathcal{E}$ for GPT-2 Medium means that the analytic target has a larger empirical MSE than the output without dropout on these masks. Monte Carlo variation and approximation error can both contribute to this difference

Model	dimension	radius CV	$\sqrt{\mathcal{V}}$	fitted RMSE	predicted $\mathcal{E}$	observed $\mathcal{E}$
GPT-2	768	1.097	7.922	8.005	0.039	0.009
GPT-2 Medium	1 024	0.739	4.629	4.605	0.015	−0.004
GPT-2 Large	1 280	0.279	2.686	2.586	0.041	0.050
GPT-2 XL	1 600	0.273	2.159	2.164	0.028	0.026
OPT-125M	768	0.391	6.947	6.931	0.362	0.383
OPT-350M	1 024	0.324	5.038	5.081	0.250	0.270
OPT-1.3B	2 048	0.170	4.878	4.861	0.139	0.159
T5 Small	512	0.757	4.641	4.828	0.240	0.296
T5 Base	768	0.495	3.662	3.691	0.050	0.019
BART Base	768	1.066	8.508	8.562	0.070	0.058

From GPT-2 to GPT-2 XL, the radius CV falls by $75.1\,\%$ and the normalized fitted RMSE falls by $73.0\,\%$. The corresponding changes are $56.6\,\%$ and $29.9\,\%$ from OPT-125M to OPT-1.3B, and $34.6\,\%$ and $23.6\,\%$ from T5 Small to T5 Base. These within-family comparisons are consistent with lower radius CV and normalized fitted RMSE at larger checkpoints. Because the checkpoints were trained separately, they do not isolate a causal effect of dimension. They also show why concentration alone does not imply a scalar-dominated regime. At native dropout, the predicted relative MSE reduction ranges from $0.015\,\%$ to $0.362\,\%$. The scalar shift remains small compared with the residual dispersion. A decrease in both terms does not by itself increase their ratio $R_{\text{shift}}/(R_{\text{shift}} + D_{\text{vis}})$.

The gain interval computed over passages is positive for every model. Relative gains range from $2.6\,\%$ to $66.5\,\%$. The difference between the two MC-1024 sets is still $17.7\,\%$ to $69.9\,\%$ of the deterministic error. The comparison is therefore made against a large finite target rather than the exact mean.

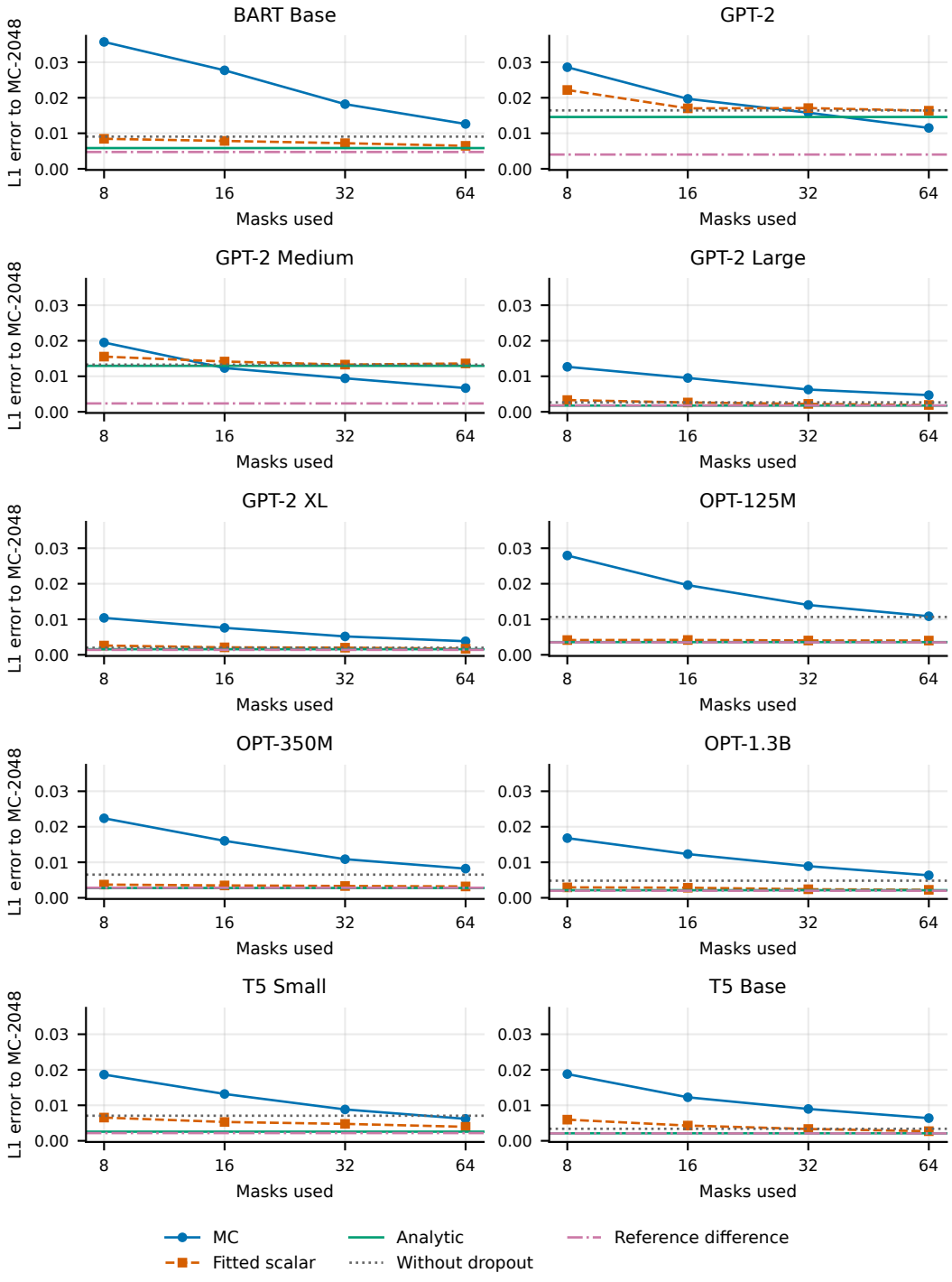


Figure 8: **Independent MC-2048 targets for all ten models.** Each panel shows the L1 error to the MC-2048 target. This target is the average of two separate MC-1024 sets. Their difference shows the remaining Monte Carlo error.

Table 12: Analytic one-pass output against the independent MC-2048 target for every language model. The interval is for the absolute L1 gain over the deterministic output. The last column gives the difference between the two MC-1024 sets as a percentage of the deterministic error.

Model	deterministic	one pass	relative gain	95 % gain interval	reference difference
GPT-2	0.01643	0.01459	11.2 %	$[0.00101, 0.00281]$	24.4 %
GPT-2 Medium	0.01329	0.01294	2.6 %	$[0.00017, 0.00055]$	17.7 %
GPT-2 Large	0.00267	0.00175	34.2 %	$[0.00067, 0.00116]$	67.0 %
GPT-2 XL	0.00199	0.00157	21.0 %	$[0.00020, 0.00065]$	69.9 %
OPT-125M	0.01065	0.00357	66.5 %	$[0.00614, 0.00808]$	32.9 %
OPT-350M	0.00652	0.00276	57.6 %	$[0.00330, 0.00422]$	43.4 %
OPT-1.3B	0.00482	0.00214	55.6 %	$[0.00208, 0.00332]$	41.4 %
T5 Small	0.00706	0.00257	63.6 %	$[0.00327, 0.00597]$	30.8 %
T5 Base	0.00338	0.00212	37.4 %	$[0.00097, 0.00157]$	61.6 %
BART Base	0.00906	0.00583	35.6 %	$[0.00284, 0.00361]$	52.2 %

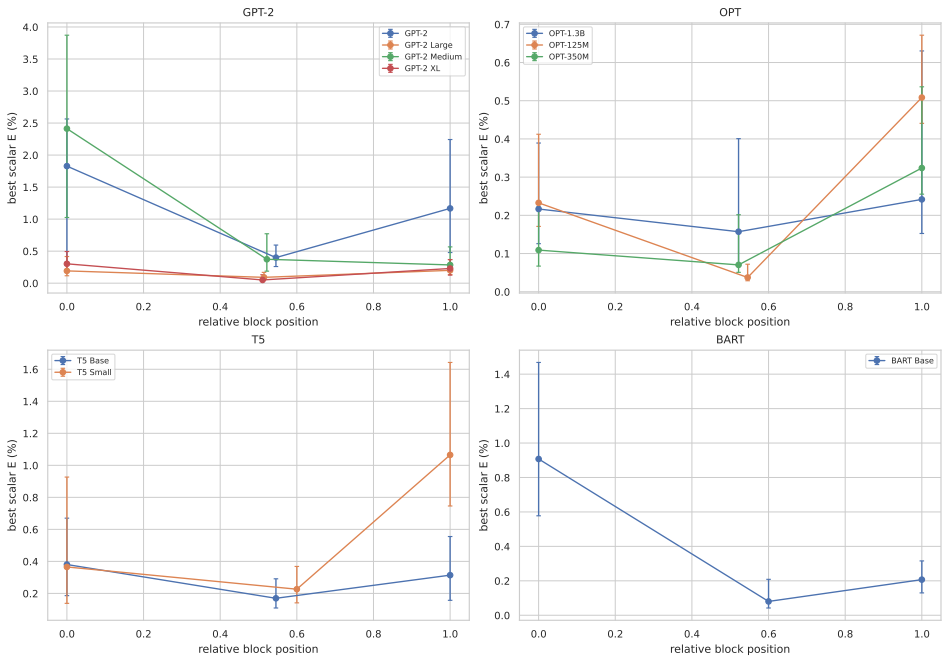


Figure 9: **Position of the masked branch.** The value is the relative MSE reduction obtained by the best fitted scalar. Bars are 95 % intervals from resampling passages.

Table 13: Relative MSE reduction from a scalar fit to the complete centered final logits after masking a branch in the first, middle, or last block. Entries are medians of reductions within passages.

Model	first	middle	last
GPT-2	1.828 %	0.399 %	1.168 %
GPT-2 Medium	2.413 %	0.373 %	0.285 %
GPT-2 Large	0.192 %	0.091 %	0.200 %
GPT-2 XL	0.302 %	0.050 %	0.229 %
OPT-125M	0.233 %	0.037 %	0.508 %
OPT-350M	0.109 %	0.070 %	0.324 %
OPT-1.3B	0.217 %	0.157 %	0.242 %
T5 Small	0.365 %	0.226 %	1.065 %
T5 Base	0.380 %	0.170 %	0.314 %
BART Base	0.908 %	0.080 %	0.207 %

For each token i, this position experiment fits $a_i \ell_{i0}$ using one coefficient shared across its masks. This rescales the complete centered logit vector, unlike the local affine family $c + a_i(\ell_{i0} - c)$. Squared errors are pooled within each passage before taking the median of the resulting reductions across passages. The largest model-and-position summary is $2.41\,\%$. Dropout in an earlier branch passes through later nonlinear blocks, so this comparison fits the final outputs rather than applying the analytic formula for the last normalized location.

Matched Gaussian noise gives results close to Bernoulli dropout in these local controls. Removing the fixed residual input raises the relative MSE reduction to about $1.9\,\%$ to $4.6\,\%$, and the prediction remains close. These observations are consistent with a leading role for second-order geometry in the tested local settings. They do not establish this conclusion for arbitrary dropout distributions or dropout throughout the network.

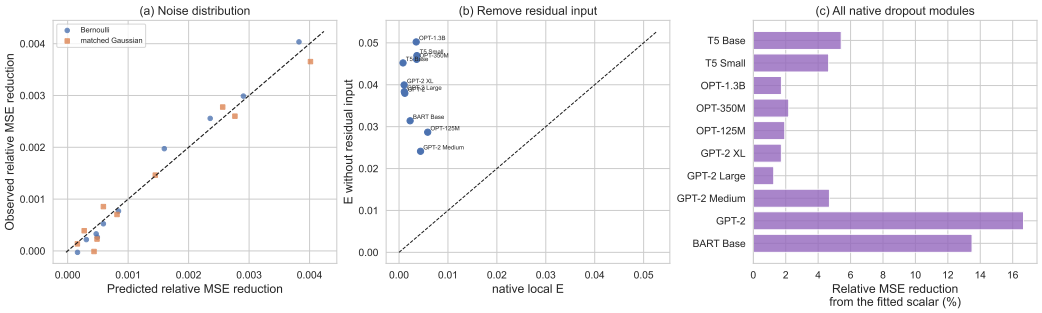


Figure 10: **Noise and residual-input comparisons for all ten models.** Gaussian noise matches the variance of Bernoulli dropout. Removing the fixed residual input makes the random branch stronger. Full-network dropout activates all native dropout operations.

Table 14: Predicted and observed relative MSE reductions in the two local comparisons. Every entry is a percentage.

	matched Gaussian		no fixed residual	
Model	predicted	observed	predicted	observed
GPT-2	0.049	0.023	3.15	3.03
GPT-2 Medium	0.016	0.014	2.48	1.93
GPT-2 Large	0.044	−0.001	3.60	3.58
GPT-2 XL	0.027	0.039	3.72	3.68
OPT-125M	0.401	0.365	2.90	2.88
OPT-350M	0.277	0.260	4.76	4.63
OPT-1.3B	0.145	0.146	3.42	3.35
T5 Small	0.256	0.278	5.28	4.33
T5 Base	0.059	0.086	3.98	4.15
BART Base	0.082	0.070	3.03	3.01

F.5 Predicting relative MSE reductions on held-out tokens

We evaluate the input-specific prediction $\mathcal{E}_{i,\mathrm{pred}} = \Lambda_i/(1+\Lambda_i)$ on the 1 200 native-rate test tokens from the ten checkpoints. Each checkpoint contributes 120 positions, with twelve positions from each of ten passages. The keep probability is $q = 0.9$, and each observed reduction $\widehat{\mathcal{E}}_i$ uses 32 masks through (92).

For each model, a constant predictor is the mean observed relative MSE reduction over its 120 native-rate validation tokens. This constant is fixed before test evaluation. A second reference predicts zero. For a prediction a_i, we average $(\widehat{\mathcal{E}}_i - a_i)^2$ over test tokens to obtain its prediction MSE. The pooled MSE uses all 1 200 tokens, giving equal weight to the ten checkpoints.

We draw the ten test passages with replacement 10 000 times. Each draw keeps all positions and checkpoints from a resampled passage together. We compute paired MSE ratios in each resample and report their 2.5th and 97.5th percentiles. Each row has its own interval, conditional on the fixed validation means.

Table 15 gives a pooled MSE reduction of $26.5\,\%$ against the validation mean, with a 95 % gain interval of $12.2\,\%$ to $37.4\,\%$. The reduction against zero is $43.2\,\%$. The ratio to the validation mean is below one for eight checkpoints, with the largest gains in the three OPT models. It ranges from 0.272 for OPT-125M to 1.495 for BART Base.

The token-level Pearson correlation is 0.609, while the correlation between the ten model medians on test is 0.973. The absolute error has mean 1.65×10^{-3} and 95th and 99th percentiles 5.53×10^{-3} and 1.25×10^{-2}. The observed reduction is negative for $30.7\,\%$ of test tokens. These prediction errors include Monte Carlo sampling variation in the 32-mask observations.

Table 15: Prediction of the relative MSE reduction on held-out test tokens. MSE is for $\mathcal{E}_{i,\mathrm{pred}}$, in units of 10^{-6}. The next two columns divide this MSE by that of the zero predictor and the model-specific validation mean. The final column gives the 95 % passage bootstrap interval for the ratio to the validation mean. Ratios below one favor the analytic prediction.

Model	MSE (10^{-6})	vs. zero	vs. validation	95 % interval
GPT-2	17.732	1.344	1.318	$[0.884, 1.667]$
GPT-2 Medium	7.654	0.998	0.992	$[0.940, 1.100]$
GPT-2 Large	2.718	0.886	0.947	$[0.862, 1.062]$
GPT-2 XL	2.378	0.820	0.868	$[0.766, 0.967]$
OPT-125M	2.241	0.091	0.272	$[0.145, 0.516]$
OPT-350M	2.293	0.172	0.381	$[0.315, 0.491]$
OPT-1.3B	0.911	0.193	0.338	$[0.224, 0.497]$
T5 Small	43.261	0.561	0.655	$[0.516, 0.961]$
T5 Base	7.161	0.674	0.676	$[0.519, 1.013]$
BART Base	4.252	1.901	1.495	$[0.831, 2.304]$
All ten models	9.060	0.568	0.735	$[0.626, 0.878]$

F.6 Additional experiments on FaceNet

FaceNet provides a fixed angular pipeline with a different geometry from both MNIST classifiers and language models. The following experiments test the law relating the mean shift to dispersion, scalar optimality, mask dependence, normalization, and the one-pass approximation on the complete LFW protocol.

In Figure 1, the six full-LFW FaceNet points use per-image medians, giving correlation 0.99998 and MAE 3.98×10^{-4}. Figure 11 instead pools squared errors across test images. For comparison with this aggregate, its predicted reduction is a weighted average of the predictions for individual images. With the empirical risk from (93), define

$$w_i := \max\{\widehat{\mathcal{Q}}_i(1), 10^{-12}\}, \qquad \mathcal{E}_{\mathrm{pred,pool}} := \frac{\sum_i w_i \mathcal{E}_{i,\mathrm{pred}}}{\sum_i w_i}.$$

The observed reduction is $1 - \sum_i \widehat{\mathcal{Q}}_i(1/T_i)/\sum_i \widehat{\mathcal{Q}}_i(1)$. Each prediction for an individual image uses no sampled mask, but these aggregation weights are estimated from 48 masks.

For the five independent Bernoulli conditions, the pooled predicted and observed relative MSE reductions have correlation above 0.99999, MAE 2.10×10^{-4}, and maximum absolute error 5.24×10^{-4}. At the native keep probability $q = 0.4$, the prediction is $1.397\,\%$ and the observed value is $1.344\,\%$. Matched Gaussian noise gives $1.397\,\%$ and $1.356\,\%$. The observed reduction is slightly below the prediction for $q = 0.4$ through 0.85 and for the Gaussian condition. The corresponding image-bootstrap intervals exclude zero, with a largest absolute difference of 0.052 percentage points. At $q = 0.95$, the interval contains zero.

At native dropout, the pooled explicit and empirical coefficients are 0.97863 and 0.97975. The risk of $1/T$ is only $0.0275\,\%$ above the best scalar fitted from 48 masks. Across test images, the direct residual ratio $\widehat{\mathcal{V}} = \widehat{\mathcal{Q}}(\hat{a})/\|(\ell_0 - c)/T\|_2^2$ has log-correlation 0.986 with $\mathcal{V}$, and their medians are 0.03168 and 0.03406. Panel (d) gives the stronger different-mask comparison described in Appendix F.2.

The dependence intervention varies the number of independent mask groups from 1 to 1 792. The median radius CV has log-log slope -0.437 ($R^2 = 0.987$). A second diagnostic compares T_i with the mean observed radius ratio for each image. For image i, write r_{i0} and r_{ij} for the normalization radii without dropout and under mask j. The radius-ratio MAE over N_{images} images is

$$\frac{1}{N_{\mathrm{images}}} \sum_{i=1}^{N_{\mathrm{images}}} \left| T_i - \frac{1}{N_{\mathrm{mask}}} \sum_{j=1}^{N_{\mathrm{mask}}} \frac{r_{ij}}{r_{i0}} \right|.$$

This diagnostic has slope -0.589 ($R^2 = 0.957$). Each slope estimates a power-law exponent in the number of independent groups, and R^2 measures how well that power law describes the sweep. These

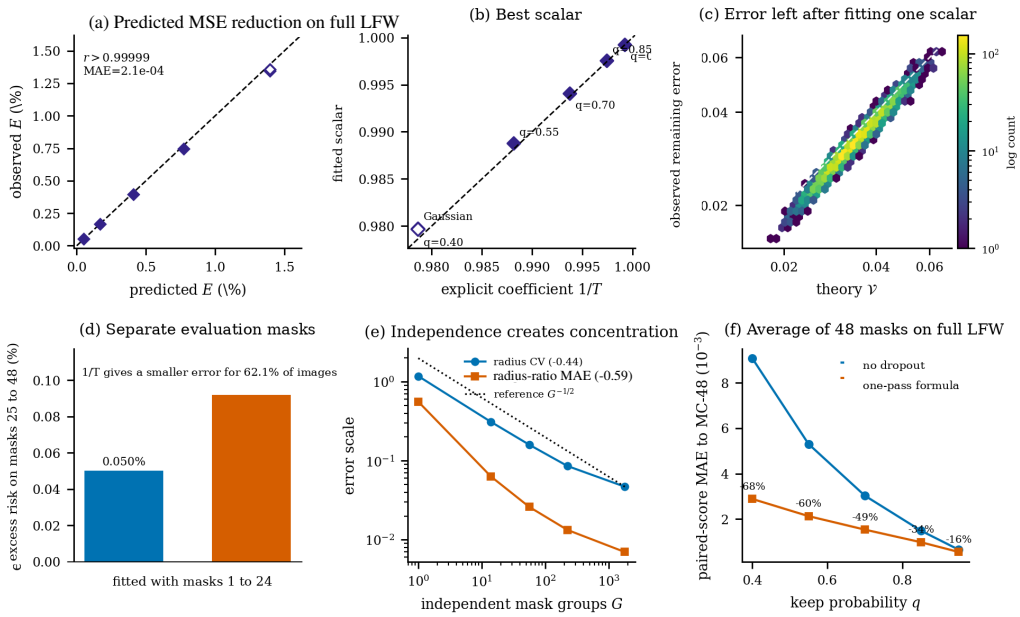


Figure 11: **Comparison of predicted and observed MSE reductions in a pretrained angular pipeline on the full LFW protocol.** The fixed FaceNet tail contains 1 792 dropout sources, an affine map to 512 dimensions, ℓ_2 normalization, and cosine scores against a fixed gallery. (a) Pooled predicted and observed MSE reductions using observed baseline-MSE weights, with intervals obtained by resampling the images. The open diamond denotes Gaussian noise. (b) Explicit and empirical scalar coefficients. (c) Direct comparison of $\mathcal{V}$ with the residual ratio. (d) Evaluation on masks different from those used by the fitted baseline. (e) Radius concentration as independent mask groups increase. (f) Approximation of MC-48 scores on all 6 000 official pairs.

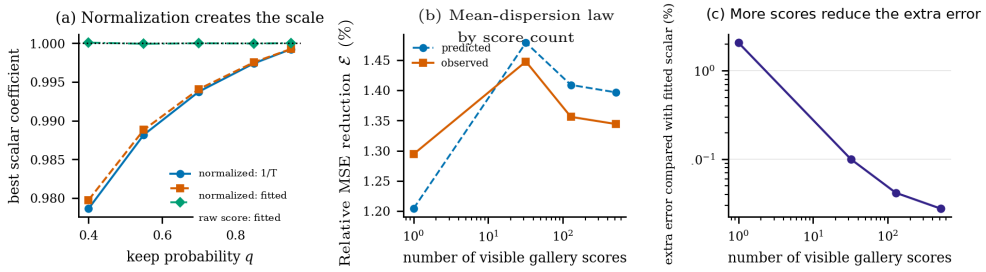


Figure 12: **Normalization and output matrix visibility isolate the mechanism.** (a) Before ℓ_2 normalization, the empirical scalar remains one, whereas after normalization it follows $1/T$. (b) The prediction from the mean shift and dispersion remains accurate as the number of visible gallery scores changes. (c) More visible gallery scores bring the explicit coefficient closer to the empirical scalar optimum.

slopes describe the tested grouping intervention, not a direct verification of the independent-replication theorem.

The scalar approximation also displays its boundary: with only 14 independent groups, the predicted relative MSE reduction is $3.54\,\%$ while the observed value is $-2.69\,\%$, and applying the predicted scalar makes the score error larger. At 224 and 1 792 groups, predicted and observed values are respectively $1.50\,\%$ versus $1.33\,\%$ and $1.39\,\%$ versus $1.35\,\%$.

The best scalar for the raw score equals 1.00008 at the original dropout rate. This confirms that the scalar contraction appears through ℓ_2 normalization. The number of visible gallery scores provides a second test. With one visible gallery score, the risk of $1/T$ is $2.08\,\%$ above the best fitted

scalar. This falls to $0.100\,\%$, $0.0416\,\%$, and $0.0275\,\%$ with 32, 128, and 512 scores. The radius is unchanged in this ablation. For the tested galleries, increasing the number of scores brings the analytic coefficient closer to the empirical scalar optimum. This change concerns the output map rather than the normalization radius.

Table 16: Approximation of the MC-48 paired cosine score. The final column is the proportion of official pairs on which the one-pass target has smaller absolute error than the original deterministic score.

q	deterministic MAE	one-pass MAE	relative gain	pair wins
0.40	0.00909	0.00288	68.3 %	79.1 %
0.55	0.00531	0.00212	59.9 %	75.6 %
0.70	0.00303	0.00153	49.5 %	72.0 %
0.85	0.00148	0.00097	34.2 %	66.9 %
0.95	0.00065	0.00054	15.9 %	60.8 %

Every uncertainty interval for the difference between the one-pass and deterministic MAE lies below zero. These intervals resample groups of scores that share an image. At the original dropout rate, ten-fold balanced accuracy under the fixed original threshold is $96.65\,\%$ for deterministic scores, $96.77\,\%$ for one-pass scores, and $96.70\,\%$ for MC-48, while all three ROC AUC values are about 0.9932. Decision agreement between one-pass and MC-48 is at least $99.9\,\%$ throughout the keep-probability sweep. The tail-only implementation is 41.5 times faster than MC-48 on 256 test images. These timings exclude the common fixed backbone and gallery products.

The single-mask relative MSE reduction and the Monte Carlo mean approximation answer different questions. Residual dispersion dominates one random pass, which makes $\mathcal{E} = 1.34\,\%$ at native dropout. Averaging 48 passes removes much of this centered dispersion while preserving the systematic scalar contraction. The analytic formula captures that contraction and therefore reduces error to the MC mean by $68.3\,\%$, even though it produces only a small relative reduction in single-pass MSE. The nearly unchanged verification accuracy is also expected: only a small number of pairs lie close enough to the decision threshold for this score correction to alter the decision.

F.6.1 Independent Monte Carlo reference for FaceNet

The two independent MC-1024 halves differ by MAE 8.83×10^{-4}. Their average gives the MC-2048 target, whose estimated sampling noise is 4.42×10^{-4}. The analytic MAE is 6.79×10^{-4}. It is therefore only 1.54 times this estimated noise level. By comparison, the deterministic score has MAE 8.40×10^{-3}.

Table 17: Error against the independent MC-2048 target. All Monte Carlo baselines use masks that do not enter the target. The final column measures variation across independent repeats.

Method	masks	repeats	mean MAE	run standard deviation
Deterministic	0	1	0.008398	n/a
Analytic one-pass	0	1	0.000679	n/a
MC-16	16	10	0.004947	0.000187
MC-48	48	10	0.002821	0.000106
MC-128	128	6	0.001773	0.000072

The analytic formula reduces MAE by $91.9\,\%$ relative to the deterministic score and wins on $92.6\,\%$ of pairs. The paired difference in absolute error is -0.00772, with image-grouped bootstrap interval $[-0.00823, -0.00718]$. This interval is conditional on the fixed MC-2048 target. A logarithmic fit over MC-16, MC-48, and MC-128 has slope -0.494, close to the usual $N_{\mathrm{MC}}^{-1/2}$ sampling rate, where N_{MC} is the number of sampled masks. The analytic error is 4.16 times smaller than MC-48 and 2.61 times smaller than MC-128. Across the three tested mask budgets, the fitted slope is consistent with the usual Monte Carlo rate. The analytic result itself uses no sampled mask. This comparison removes the main ambiguity of using MC-48 as both a target and a baseline.

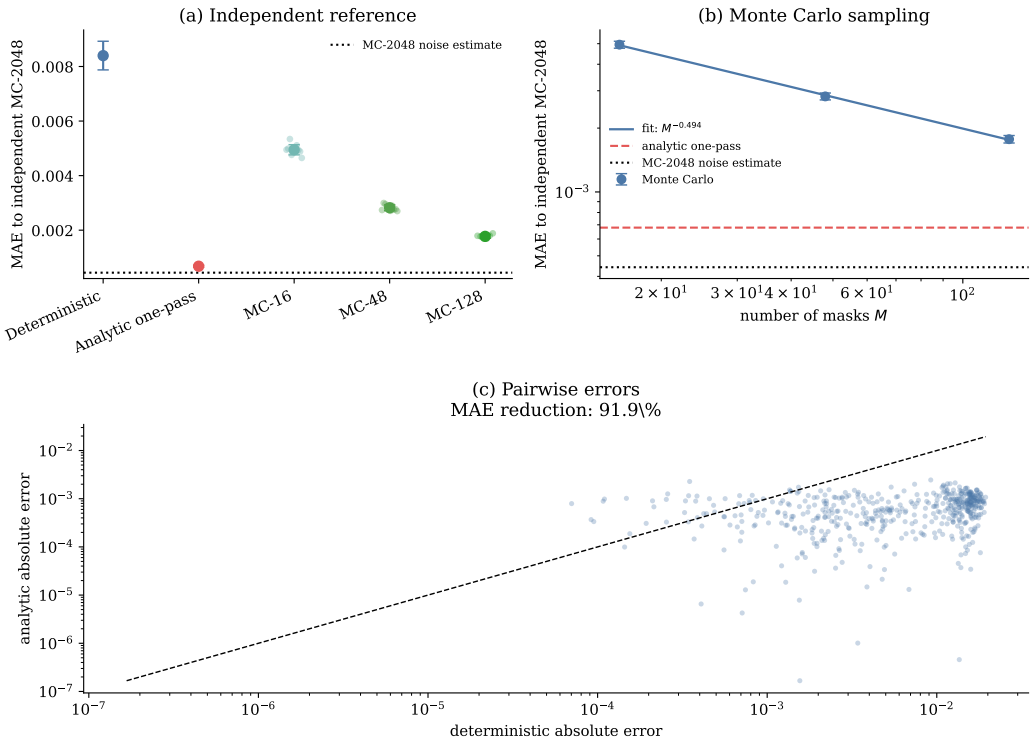


Figure 13: **FaceNet against an independent MC-2048 target.** The experiment uses 512 official LFW pairs balanced by identity class and fold. (a) The analytic one-pass target has lower MAE than all three sampled Monte Carlo baselines tested. Points show independent repeats and bars show paired bootstrap intervals. (b) Monte Carlo error follows $N_{\mathrm{MC}}^{-0.494}$, while the analytic result uses no sampled mask. (c) Each point compares analytic and deterministic absolute errors on one pair. The diagonal marks equal error.

F.7 Further MNIST and language model boundary tests

Small MNIST models. The probability-space comparison below complements the predicted and observed MSE reductions in logit space by asking whether one analytic pass also approximates the average class probabilities from repeated dropout passes.

Table 18: One-pass probability approximation on the small MNIST models. Values are means over the first 512 test images and three seeds at the widest architecture, with $q = 0.65$. Top-class agreement compares the analytic prediction with the mean of 20 sampled probability vectors.

Architecture	ℓ_1 distance to MC-20	top-class agreement
Residual MLP	0.00138	100.00 %
Mini-ViT	0.00396	99.61 %
PreAct ResNet	0.00169	100.00 %

The two independent ten-mask halves differ by mean L1 distances of 0.00222, 0.00603, and 0.00306, respectively. MC-20 is therefore a finite sampled reference, not the exact mean. Agreement with this reference does not imply improved classification or calibration. For example, mean ResNet NLL on this subset increases from 0.0523 without dropout to 0.0568 with the analytic correction, compared with 0.0573 for MC-20.

Complete language models. Table 12 gives the corresponding language model comparison against independent MC-2048 targets. We also activate the native dropout operations throughout each complete language model. This gives a separate test of the boundary between the local theorem and

the full network. For each token i with $\ell_{i0} \neq 0$, the fitted coefficient is shared across its 16 masks:

$$a_i := \arg\min_{a \in \mathbb{R}} \sum_{j=1}^{16} \|\ell_{ij} - a\ell_{i0}\|_2^2.$$

This diagnostic rescales the complete centered logit vector, unlike the local affine family, which keeps c fixed. Squared errors are pooled within each passage, and the table reports the median of these reductions across passages relative to the output without dropout.

Table 19: Full-network boundary test on all ten language models. Native dropout is active throughout each model. The local column uses the analytic affine target. The column for dropout throughout the network uses the complete-logit scalar fit defined above. These are different approximation families.

Model	local predicted $\mathcal{E}$	full network fitted $\mathcal{E}$	top-token change rate
GPT-2	0.039 %	16.67 %	32.8 %
GPT-2 Medium	0.015 %	4.70 %	31.2 %
GPT-2 Large	0.041 %	1.26 %	10.9 %
GPT-2 XL	0.028 %	1.73 %	10.9 %
OPT-125M	0.362 %	1.94 %	17.2 %
OPT-350M	0.250 %	2.19 %	21.9 %
OPT-1.3B	0.139 %	1.73 %	7.8 %
T5 Small	0.240 %	4.65 %	45.3 %
T5 Base	0.050 %	5.42 %	29.7 %
BART Base	0.070 %	13.49 %	62.5 %

The largest model-level median reduction is $16.7\,\%$. This is a summary of the tested models, not a bound for every token or network. With dropout throughout the network, most of the measured squared change remains after this scalar fit.