Answer Bandwidth: Why Sigmoid Fails in Hidden Layers

Jaepil Jeong
Cognica, Inc.
Email: jaepil@cognica.io
Date: March 27, 2026

"The theory of probabilities is at bottom nothing but common sense reduced to calculus."
— Pierre-Simon Laplace, *Theorie analytique des probabilites*, 1812

Abstract

The standard explanation for why sigmoid activations fail in hidden layers of deep networks is *vanishing gradients*: repeated multiplication by $\sigma'(z) \in (0, 0.25]$ attenuates error signals exponentially with depth. We present experimental evidence that this explanation is incomplete and identify a more fundamental mechanism: **answer bandwidth compression**. Sigmoid's bounded output range $(0, 1)$ reduces the effective rank of hidden representations, destroying representational capacity independently of gradient flow. We establish this through a six-experiment falsification-refinement arc on CIFAR-10 (VGG-style CNN) and WikiText-2 (GPT-2 Transformer), testing 60+ configurations across 30 hypotheses.

The arc proceeds as follows. Experiment 1 (seq) **refutes** the compositional "question sequencing" prediction from Jeong (2026) — that ordering activation functions by their probabilistic semantics improves performance — establishing that the dominant variable is not question semantics but representational capacity. Experiment 2 (bw) **validates** the refined hypothesis: sigmoid accuracy is strictly monotonic with temperature β (88.17% at $\beta = 0.25$ to 89.40% at $\beta = 4.0$), while Swish $x \cdot \sigma(\beta x)$ is approximately β-invariant (range 1.30%), confirming that the multiplicative x factor preserves "answer bandwidth" regardless of gating strength. Experiment 3 (gate) confirms that the gate function is secondary: GELU $x \cdot \Phi(\beta x)$ and Swish $x \cdot \sigma(\beta x)$ are experimentally indistinguishable at matched β (max $|\Delta| = 0.44\%$), with all 9 configurations falling on a single rank-accuracy curve (Pearson $r = 0.94$, RMSE 0.13%). Experiment 4 (bypass) establishes a sharp scope boundary: additive skip connections do NOT rescue sigmoid (mean rescue $= -0.06\%$) because compression occurs *inside* the block before the shortcut can intervene — answer bandwidth is an intra-activation property of the x factor, not an inter-block pathway property. Experiment 5 (depth) resolves the depth dimension: the sigmoid penalty *decreases* with depth (2.45% at D=1, 0.64% at D=2, 0.30% at D=3), directly contradicting the vanishing gradient prediction while confirming that bandwidth compression is the primary mechanism modulated by network capacity. Experiment 6 (xfer) demonstrates cross-architecture transfer to Transformers: sigmoid perplexity is monotonically decreasing with β in GPT-2 FFN layers (80.27 to 58.23), but Swish β-invariance breaks (PPL range 16.98) because the Transformer's structural residual stream provides partial bandwidth compensation absent in CNNs.

We conclude with a five-level hierarchy of bandwidth preservation: (1) the gate function (σ vs Φ) is freely interchangeable, (2) only multiplicative bypass (x factor) preserves bandwidth — additive bypass (skip connections) cannot rescue it, (3) bandwidth must be preserved within the activation (intra-activation), not around the block (inter-block), (4) the penalty is depth-modulated but not depth-dependent in the vanishing gradient sense, and (5) the mechanism transfers across architectures with differing mediating pathways.

1. Introduction

1.1 The Sigmoid Hidden Layer Problem

Every deep learning textbook answers the question "why not use sigmoid in hidden layers?" with the same explanation: vanishing gradients. The sigmoid derivative $\sigma'(z) = \sigma(z)(1 - \sigma(z))$ has a maximum of 0.25 at $z = 0$ and decays exponentially as $|z|$ grows. In a network of depth L, the gradient signal at layer l is attenuated by a factor of $\prod_{k=l+1}^{L} \sigma'(z_k) \leq 0.25^{L-l}$, making early layers untrainable (Hochreiter, 1991; Bengio et al., 1994; Glorot & Bengio, 2010).

This explanation is widely accepted and practically useful — it motivated the adoption of ReLU (Nair & Hinton, 2010) and its descendants. Yet it conflates two distinct mechanisms:

1. **Gradient attenuation**: the backward-pass signal weakens with depth.
2. **Representational compression**: the forward-pass representation loses information.

Both mechanisms produce the same symptom (poor training with sigmoid in hidden layers), but they make different predictions about how the penalty scales with depth, how bypass mechanisms (skip connections, the x factor in Swish) mitigate it, and which architectural interventions are effective. Disentangling them requires controlled experiments that the original vanishing gradient analyses did not perform.

1.2 The Question Sequencing Prediction

In the companion paper (Jeong, 2026), the correspondence between activation functions and probabilistic questions is established:

- **Sigmoid**: "How probable is the hypothesis?" — posterior probability in $(0, 1)$ (Theorem 6.4.1)
- **ReLU**: "How much evidence is present?" — MAP estimate in $[0, +\infty)$ (Theorem 6.5.3)
- **Swish**: "What is the expected relevant signal?" — Bayesian expected value (Theorem 6.7.4)
- **GELU**: Gaussian (probit) approximation of Swish (Theorem 6.8.1)

Section 9.2 of Jeong (2026) extends this to predict that *ordering* these questions optimally — magnitude detection (ReLU) in lower layers, probability estimation (sigmoid) in upper layers — should outperform the reverse, because feature extraction precedes decision-making in any reasonable inference pipeline.

This paper tests and refutes this compositional prediction, identifies the true mechanism (answer bandwidth compression), and traces it through five subsequent experiments that establish its scope, boundaries, and cross-architecture generality.

1.3 The Present Contribution

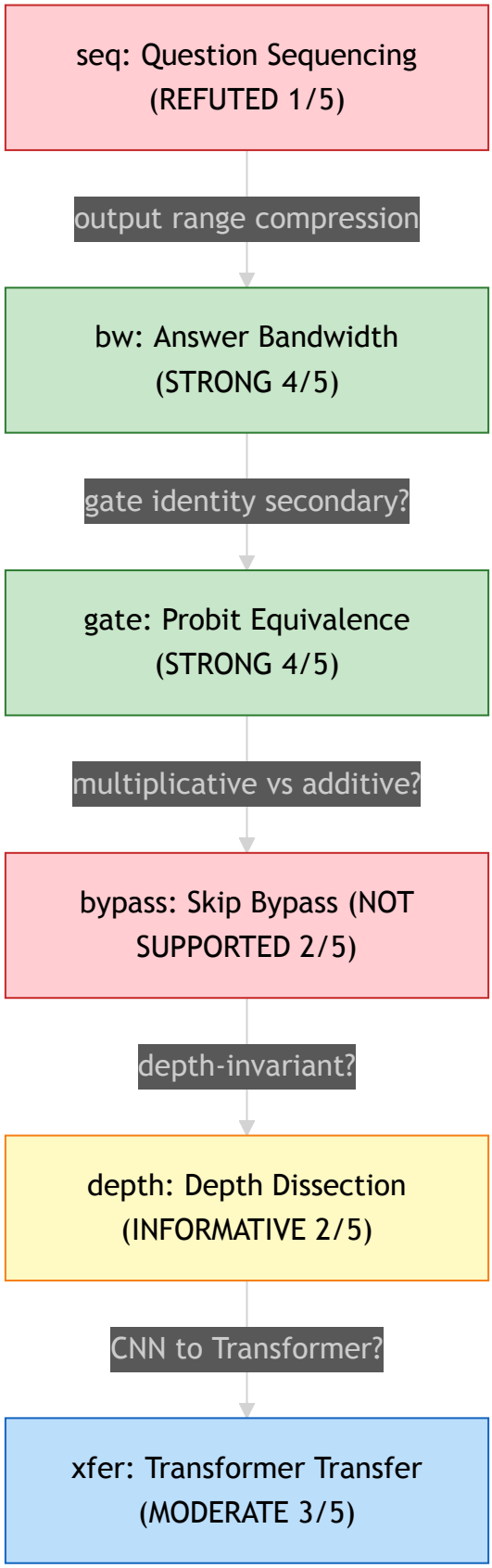
We establish the following results through six controlled experiments:

1. **Falsification of question sequencing** (Section 3): The compositional prediction fails decisively. The "wrong" ordering (sigmoid$\rightarrow$ReLU) outperforms the "right" ordering (ReLU $\rightarrow$sigmoid) by 1.48%. Only 1 of 5 hypotheses is supported. The dominant variable is not question semantics but activation output range.
2. **The answer bandwidth mechanism** (Section 4): Sigmoid fails because its output range

$(0, 1)$ compresses effective rank. Temperature-scaled sigmoid $\sigma(\beta x)$ shows strictly monotonic accuracy with β, while generalized Swish $x \cdot \sigma(\beta x)$ is approximately β-invariant. Effective rank mediates both spectra ($r = 0.82$, RMSE 0.68%). This is strong support (4/5 hypotheses).

3. **Gate function interchangeability** (Section 5): GELU and Swish are experimentally indistinguishable at matched β (max $|\Delta| = 0.44\%$), with the 1.702x probit scaling confirmed quantitatively. The rank-accuracy curve tightens to $r = 0.94$, RMSE 0.13% when both spectra preserve bandwidth.
4. **Additive bypass failure** (Section 6): Skip connections cannot rescue sigmoid (mean rescue $= -0.06\%$). Compression occurs inside the block before the shortcut can intervene. Only multiplicative bypass (the x factor) preserves bandwidth at the element level.
5. **Depth dissection** (Section 7): The sigmoid penalty *decreases* with depth (2.45%$\rightarrow$0.64%$\rightarrow$ 0.30%), contradicting the vanishing gradient prediction. Bandwidth compression is the primary mechanism, modulated by network capacity. Swish is depth-invariant (max penalty 0.33%).
6. **Cross-architecture transfer** (Section 8): The bandwidth hypothesis transfers from CNNs to Transformers. Sigmoid perplexity decreases monotonically with β in GPT-2. The Transformer residual stream provides genuine bandwidth dampening (ratio = 0.27), unlike CNN skip connections (ratio ≈ 0).

Figure 1.1 (Falsification-Refinement Arc: seq through xfer).



> Red = refuted/not supported, green = strong support, yellow = mixed, blue = cross-architecture transfer. Each experiment answers the question raised by its predecessor.

1.4 Notation

Symbol	Definition
$\sigma(x)$	Sigmoid function: $1/(1+\exp(-x))$
$\Phi(x)$	Standard Gaussian CDF
$\sigma_\beta(x)$	Temperature-scaled sigmoid: $\sigma(\beta x)$
$\text{Swish}_\beta(x)$	Generalized Swish: $x \cdot \sigma(\beta x)$
$\text{GELU}_\beta(x)$	Generalized GELU: $x \cdot \Phi(\beta x)$
$r_{\text{eff}}(A)$	Effective rank of matrix A (Definition 2.2.1)
D	Number of convolutional blocks per stage
PPL	Perplexity: $\exp(\text{CE loss})$

2. Mathematical Preliminaries

2.1 The Generalized Swish Spectrum

Definition 2.1.1 (Generalized Swish). For $\beta > 0$, the generalized Swish function is:

$$\text{Swish}_\beta(x) = x \cdot \sigma(\beta x) \tag{1}$$

Theorem 2.1.2 (Certainty Spectrum; Jeong, 2026, Theorem 6.7.6). The generalized Swish parametrizes a continuous spectrum between maximum ignorance, Bayesian estimation, and deterministic thresholding:

$$\lim_{\beta \to 0} x \cdot \sigma(\beta x) = \frac{x}{2} \quad (\text{uniform prior: maximum ignorance}) \tag{2}$$

$$\beta = 1: \quad x \cdot \sigma(x) = \text{Swish}(x) \quad (\text{canonical Bayesian posterior}) \tag{3}$$

$$\lim_{\beta \to \infty} x \cdot \sigma(\beta x) = \text{ReLU}(x) \quad (\text{deterministic MAP}) \tag{4}$$

Definition 2.1.3 (Generalized Sigmoid Spectrum). For $\beta > 0$, the temperature-scaled sigmoid is:

$$\sigma_\beta(x) = \sigma(\beta x) = \frac{1}{1 + \exp(-\beta x)} \tag{5}$$

Proposition 2.1.4 (Output Range Distinction). The generalized sigmoid and generalized Swish differ fundamentally in their output ranges:

(i) $\sigma_\beta : \mathbb{R} \to (0, 1)$ for all $\beta > 0$. The output is bounded regardless of input magnitude.

(ii) $\text{Swish}_\beta : \mathbb{R} \to [\approx -0.278/\beta, +\infty)$. The output is unbounded above, scaling linearly with input magnitude for large $|x|$.

Proof. For (i): $0 < \sigma(\beta x) < 1$ for all $x \in \mathbb{R}$ and $\beta > 0$, by the range of the sigmoid. For (ii): as $x \to +\infty$, $\sigma(\beta x) \to 1$, so $x \cdot \sigma(\beta x) \to x$, which is unbounded. The lower bound $\approx -0.278/\beta$ is the global minimum of $x \cdot \sigma(\beta x)$, found by setting $\frac{d}{dx}[x \cdot \sigma(\beta x)] = 0$. $\square$

Remark 2.1.5 This output range distinction is the core of the answer bandwidth hypothesis. The x factor in $\text{Swish}_\beta(x) = x \cdot \sigma(\beta x)$ serves as a multiplicative bypass that preserves input magnitude information. The gate $\sigma(\beta x)$ modulates the signal but never confines it to a fixed interval. In contrast, $\sigma_\beta(x)$ maps all inputs to $(0, 1)$, discarding magnitude information entirely.

2.2 Effective Rank

Definition 2.2.1 (Effective Rank; Roy & Vetterli, 2007). For a matrix A with singular values $s_1 \geq s_2 \geq \cdots \geq s_n > 0$, the effective rank is:

$$r_{\text{eff}}(A) = \exp\left(-\sum_{i=1}^{n} p_i \log p_i \right) \tag{6}$$

where $p_i = s_i / \sum_j s_j$ is the normalized singular value distribution.

Proposition 2.2.2 (Effective Rank Properties).

(i) $1 \leq r_{\text{eff}}(A) \leq \text{rank}(A) \leq \min(m, n)$ for $A \in \mathbb{R}^{m \times n}$.

(ii) $r_{\text{eff}}(A) = \text{rank}(A)$ if and only if all non-zero singular values are equal (uniform spectrum).

(iii) $r_{\text{eff}}(A) = 1$ if and only if $\text{rank}(A) = 1$ (all energy concentrated in one singular value).

Proof. This follows from the properties of Shannon entropy applied to the normalized singular value distribution. The entropy $H(p) = -\sum p_i \log p_i$ satisfies $0 \leq H(p) \leq \log n$, with equality on the right iff $p_i = 1/n$ for all i. □

Remark 2.2.3 Effective rank captures the "usable dimensionality" of a representation. If a 256-dimensional hidden layer has effective rank 100, approximately 100 dimensions carry meaningful variation while the remaining 156 are redundant or near-zero. The answer bandwidth hypothesis predicts that bounded activations reduce effective rank by compressing the output range, while the x factor preserves it.

2.3 The Probit Approximation

Proposition 2.3.1 (Probit-Logistic Scaling; Jeong, 2026, Proposition 6.8.2). The Gaussian CDF $\Phi(x)$ and the logistic sigmoid $\sigma(x)$ are related by:

$$\Phi(x) \approx \sigma(1.702\, x) \tag{7}$$

Therefore, the generalized GELU at temperature β approximates the generalized Swish at temperature $1.702\,\beta$:

$$\text{GELU}_\beta(x) = x \cdot \Phi(\beta x) \approx x \cdot \sigma(1.702\,\beta\, x) = \text{Swish}_{1.702\beta}(x) \tag{8}$$

---

3. Experiment 1: The Question Sequencing Hypothesis (seq)

3.1 Experimental Setup

Definition 3.1.1 (Architecture). A VGG-style CNN with 3 convolutional blocks and approximately 1.15M parameters:

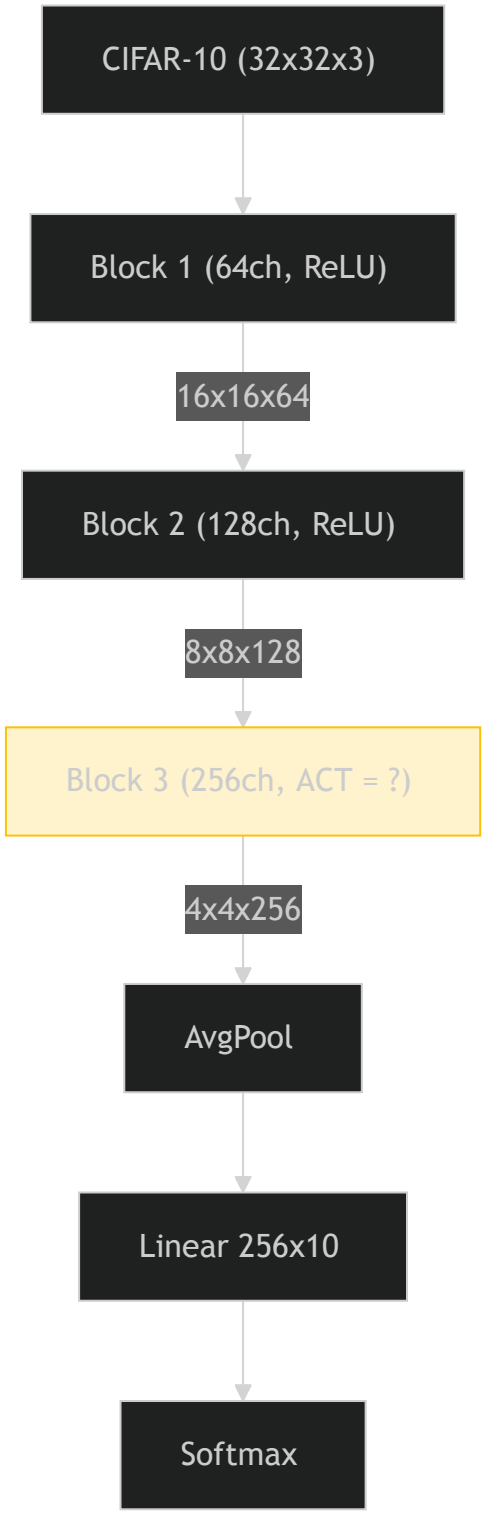
- Block 1: [Conv(3,64)-BN-Act-Conv(64,64)-BN-Act-MaxPool] $\to 16 \times 16 \times 64$

- Block 2: [Conv(64,128)-BN-Act-Conv(128,128)-BN-Act-MaxPool] $\rightarrow 8 \times 8 \times 128$
- Block 3: [Conv(128,256)-BN-Act-Conv(256,256)-BN-Act-MaxPool] $\rightarrow 4 \times 4 \times 256$
- Output: AdaptiveAvgPool2d(1) $\rightarrow$ Flatten(256) $\rightarrow$ Linear(256, 10)

Training: 50 epochs, batch size 128, Adam (lr=10^{-3}), CIFAR-10 with standard augmentation (RandomCrop, HorizontalFlip).

Figure 3.1 (VGG-style CNN Architecture).

Each block repeats: Conv $\rightarrow$ BN $\rightarrow$ Act $\rightarrow$ Conv $\rightarrow$ BN $\rightarrow$ Act $\rightarrow$ MaxPool.



> Block 3 activation (highlighted) is the experimental variable. Blocks 1--2 use ReLU in all experiments except seq. Each block internally consists of Conv-BN-Act-Conv-BN-Act-MaxPool.

Definition 3.1.2 (Configurations). Eight activation configurations, where [b1, b2, b3] denotes the activation function for each block:

Config	[Block1, Block2, Block3]	Rationale
seq-R	[ReLU, ReLU, ReLU]	Uniform baseline
seq-S	[sigmoid, sigmoid, sigmoid]	Uniform baseline
seq-G	[GELU, GELU, GELU]	Uniform baseline
seq-W	[Swish, Swish, Swish]	Uniform baseline
seq-RS	[ReLU, ReLU, sigmoid]	Framework-optimal: "how much?" $\rightarrow$ "how probable?"
seq-SR	[sigmoid, sigmoid, ReLU]	Reversed: "how probable?" $\rightarrow$ "how much?"
seq-GS	[GELU, GELU, sigmoid]	Expected value $\rightarrow$ probability
seq-RG	[ReLU, ReLU, GELU]	MAP $\rightarrow$ Bayes estimate

3.2 Hypotheses

The question sequencing framework (Jeong, 2026, Section 9.2) predicts:

Hypothesis 3.2.1 (H1, Critical): $\mathrm{Acc}(\text{seq-RS}) > \mathrm{Acc}(\text{seq-SR})$. The framework-optimal ordering ("how much?" in lower layers, "how probable?" in upper layers) should outperform the reversed ordering.

Hypothesis 3.2.2 (H2): $\mathrm{Acc}(\text{seq-RS}) \geq \mathrm{Acc}(\text{seq-R})$. Adding sigmoid at the top should help or at least not hurt.

Hypothesis 3.2.3 (H3): $\mathrm{Acc}(\text{seq-GS}) \geq \mathrm{Acc}(\text{seq-G})$. Sigmoid on top of GELU should help.

Hypothesis 3.2.4 (H4): seq-SR is the worst of $\{$R, S, RS, SR$\}$. The reversed ordering should be worst.

Hypothesis 3.2.5 (H5): $\mathrm{Acc}(\text{seq-RG}) \geq \mathrm{Acc}(\text{seq-R})$. Complementary questions should help.

3.3 Results

Config	Accuracy	Effective Rank (Block 3)	L0 Sparsity	ECE
seq-R	91.00%	215	0.30	0.041
seq-S	87.49%	149	0.00	0.027
seq-G	91.22%	217	0.01	0.038
seq-W	91.30%	218	0.01	0.039
seq-RS	88.55%	152	--	--
seq-SR	90.03%	207	--	--
seq-GS	88.82%	149	--	--
seq-RG	91.42%	219	--	--

Theorem 3.3.1 (Refutation of Question Sequencing). The critical test H1 fails decisively:

$$\text{Acc}(\text{seq-SR}) - \text{Acc}(\text{seq-RS}) = 90.03\% - 88.55\% = +1.48\% \tag{9}$$

The "wrong" ordering outperforms the "right" ordering. Only H5 is supported (seq-RG $\geq$ seq-R: 91.42% vs 91.00%, a near-trivial result since GELU $\approx$ ReLU for positive inputs).

Hypothesis	Prediction	Observed	Verdict
H1: RS > SR	RS best	SR > RS by 1.48%	**REFUTED**
H2: RS $\geq$ R	RS helps	RS < R by 2.45%	**REFUTED**
H3: GS $\geq$ G	GS helps	GS < G by 2.40%	**REFUTED**
H4: SR worst	SR worst	S worst (87.49%)	**REFUTED**
H5: RG $\geq$ R	RG helps	RG > R by 0.42%	**SUPPORTED**

3.4 The Compression Mechanism

Observation 3.4.1 (Effective Rank Pattern). Every configuration with sigmoid in block 3 shows dramatic rank reduction:

Config	Block 3 Activation	Effective Rank	Accuracy
seq-R	ReLU	215	91.00%
seq-SR	ReLU	207	90.03%
seq-RG	GELU	219	91.42%
seq-RS	sigmoid	152	88.55%
seq-GS	sigmoid	149	88.82%
seq-S	sigmoid	149	87.49%

The rank reduction is approximately 60--70 units whenever sigmoid appears in block 3, regardless of what precedes it. The accuracy penalty correlates with this rank reduction, not with question ordering.

Remark 3.4.2 (BatchNorm Compensation). Sigmoid blocks develop elevated BatchNorm gamma values (up to 2.6x baseline), attempting to compensate for the compressed output range. This compensation is insufficient: the BN affine transform can scale and shift but cannot restore the singular value diversity lost by mapping all outputs to $(0, 1)$.

Remark 3.4.3 (Descriptive Validity). The "question" characterizations remain descriptively valid: ReLU produces the sparsest activations (L0=0.30), sigmoid produces the best-calibrated predictions (ECE=0.027), and GELU/Swish occupy intermediate positions. However, these properties do not compose sequentially as the framework predicts. The failure is specifically the *compositional* claim, not the individual characterizations.

---

4. Experiment 2: The Answer Bandwidth Hypothesis (bw)

4.1 From Refutation to Refinement

Experiment 1 refuted the question sequencing prediction but identified a clear mechanism: sigmoid's bounded output range $(0, 1)$ reduces effective rank. This motivates a refined hypothesis.

Definition 4.1.1 (Answer Bandwidth). The *answer bandwidth* of an activation function f is the effective range of its output distribution when applied to a typical pre-activation distribution. Formally, for pre-activations $z \sim \mathcal{D}$, the answer bandwidth is the interquartile range $\mathrm{IQR}(f(z))$.

Hypothesis 4.1.2 (Answer Bandwidth Hypothesis). Sigmoid fails in hidden layers not because "how probable?" is the wrong question (refuted in Experiment 1), but because confining the answer to $(0, 1)$ creates an information bottleneck. The x factor in Swish ($x \cdot \sigma(\beta x)$) prevents this bottleneck by preserving output range proportional to input magnitude.

4.2 Experimental Setup

The same VGG architecture as Experiment 1 (Definition 3.1.1). Only block 3's activation varies; blocks 1--2 are always ReLU. This isolates the bandwidth effect from cross-block interaction.

Definition 4.2.1 (Configurations). Ten configurations spanning two activation spectra:

Config	Block 3 Activation	Spectrum
bw-R	ReLU	Baseline
bw-sig025	$\sigma(0.25x)$	Sigmoid, $\beta = 0.25$
bw-sig050	$\sigma(0.5x)$	Sigmoid, $\beta = 0.5$
bw-sig100	$\sigma(x)$	Sigmoid, $\beta = 1.0$
bw-sig200	$\sigma(2x)$	Sigmoid, $\beta = 2.0$
bw-sig400	$\sigma(4x)$	Sigmoid, $\beta = 4.0$
bw-swi025	$x \cdot \sigma(0.25x)$	Swish, $\beta = 0.25$
bw-swi050	$x \cdot \sigma(0.5x)$	Swish, $\beta = 0.5$
bw-swi100	$x \cdot \sigma(x)$	Swish, $\beta = 1.0$
bw-swi200	$x \cdot \sigma(2x)$	Swish, $\beta = 2.0$

4.3 Hypotheses

Hypothesis 4.3.1 (H1): Sigmoid accuracy is monotonically non-decreasing with β. Higher β widens the effective output range within $(0, 1)$.

Hypothesis 4.3.2 (H2): Swish accuracy is approximately β-invariant (range $< 2\%$). The x factor preserves bandwidth regardless of gating strength.

Hypothesis 4.3.3 (H3): Block 3 effective rank increases with β for the sigmoid spectrum.

Hypothesis 4.3.4 (H4): Block 3 effective rank is approximately β-invariant for the Swish spectrum (range < 30).

Hypothesis 4.3.5 (H5): Accuracy vs effective rank forms a single curve across both spectra (RMSE $< 2\%$).

4.4 Results

Theorem 4.4.1 (Sigmoid Monotonicity). Sigmoid accuracy is strictly monotonic with β:

Config	β	Accuracy	Effective Rank	BN2 Gamma
bw-sig025	0.25	88.17%	108	4.05
bw-sig050	0.50	88.28%	137	3.21
bw-sig100	1.00	88.55%	152	2.45
bw-sig200	2.00	89.23%	183	2.08
bw-sig400	4.00	89.40%	210	1.79

Interpretation. As β increases, $\sigma(\beta x)$ transitions from a near-constant function (output ≈ 0.5 for all x) toward a step function (output ≈ 0 or ≈ 1). The effective output range widens: at $\beta = 0.25$, the IQR is approximately 0.35; at $\beta = 4.0$, it is approximately 0.91. Wider effective range $\implies$ higher effective rank $\implies$ higher accuracy.

Theorem 4.4.2 (Swish Invariance). Swish accuracy is approximately β-invariant:

Config	β	Accuracy	Effective Rank	BN2 Gamma
bw-swi025	0.25	90.21%	164.7	1.47
bw-swi050	0.50	90.78%	186.2	1.42
bw-swi100	1.00	91.51%	205.8	1.38
bw-swi200	2.00	90.94%	216.1	1.33
bw-R (ReLU)	∞	91.00%	215	1.35

Accuracy range: $91.51\% - 90.21\% = 1.30\%$, well within the 2% invariance threshold. All Swish configurations are within 0.8% of the ReLU baseline.

Theorem 4.4.3 (Rank-Accuracy Mediation). Effective rank mediates both sigmoid and Swish accuracy across all 10 configurations:

$$\text{Pearson } r = 0.82, \quad \text{RMSE} = 0.68\% \tag{10}$$

All 10 data points — 5 sigmoid and 5 Swish — fall on a single rank-accuracy curve, demonstrating that effective rank is the common mediating variable regardless of whether the activation is bounded or unbounded.

Theorem 4.4.4 (BatchNorm Compensation Pattern). The BN2 gamma statistic reveals the compensation mechanism:

- Sigmoid: BN2 gamma decreases monotonically from 4.05 ($\beta = 0.25$, maximum compression $\implies$ maximum compensation) to 1.79 ($\beta = 4.0$, near-step function $\implies$ minimal compensation).
- Swish: BN2 gamma remains stable at 1.33--1.47, near the ReLU baseline of 1.35.

Interpretation. BatchNorm's affine transform $\gamma \cdot \hat{x} + \beta_{\text{BN}}$ attempts to rescale the compressed sigmoid outputs. The high gamma values at low β indicate extreme compensation effort, but this linear rescaling cannot restore the singular value diversity destroyed by the nonlinear compression.

4.5 Hypothesis Verdicts

Hypothesis	Criterion	Observed	Verdict
H1: Sigmoid monotonic	Acc increases with β	88.17% $\rightarrow$ 89.40%, strictly monotonic	**SUPPORTED**
H2: Swish invariant	Acc range $< 2\%$	Range = 1.30%	**SUPPORTED**
H3: Sigmoid rank increases	Rank increases with β	108 $\rightarrow$ 210, strictly monotonic	**SUPPORTED**
H4: Swish rank invariant	Rank range < 30	Range = 51.4	**REFUTED**
H5: Single rank-acc curve	RMSE $< 2\%$	RMSE = 0.68%, $r = 0.82$	**SUPPORTED**

Overall: 4/5 hypotheses supported — STRONG SUPPORT.

Remark 4.5.1 (H4 Refutation is Informative). Swish effective rank varies substantially (164.7 to 216.1), yet accuracy varies by only 1.30%. This indicates that the x factor preserves accuracy through a mechanism beyond effective rank alone — possibly gradient flow quality. Block 3 gradient norms are consistently higher for Swish (0.16–0.42) than for low-β sigmoid (0.07–0.12), suggesting the x factor maintains both representational capacity and gradient signal quality.

---

5. Experiment 3: Gate Function Interchangeability (gate)

5.1 Motivation

Experiment 2 established that the x factor preserves answer bandwidth. A natural question follows: does the identity of the gate function (σ vs Φ) matter? The probit approximation (Proposition 2.3.1) predicts it should not, since $\Phi(x) \approx \sigma(1.702x)$.

5.2 Experimental Setup

The same VGG architecture, with block 3 varied across 9 configurations:

Config	Block 3 Activation	Spectrum
gate-R	ReLU	Baseline
gate-swi025	$x \cdot \sigma(0.25x)$	Swish, $\beta = 0.25$
gate-swi050	$x \cdot \sigma(0.5x)$	Swish, $\beta = 0.50$
gate-swi100	$x \cdot \sigma(x)$	Swish, $\beta = 1.00$
gate-swi200	$x \cdot \sigma(2x)$	Swish, $\beta = 2.00$
gate-gel025	$x \cdot \Phi(0.25x)$	GELU, $\beta = 0.25$
gate-gel050	$x \cdot \Phi(0.5x)$	GELU, $\beta = 0.50$
gate-gel100	$x \cdot \Phi(x)$	GELU, $\beta = 1.00$
gate-gel200	$x \cdot \Phi(2x)$	GELU, $\beta = 2.00$

5.3 Hypotheses

Hypothesis 5.3.1 (H1): GELU accuracy is approximately β-invariant (range $< 2\%$).

Hypothesis 5.3.2 (H2): At matched β, $|\text{Acc}(\text{GELU}_\beta) - \text{Acc}(\text{Swish}_\beta)| < 1.0\%$.

Hypothesis 5.3.3 (H3): At matched β, $|r_{\text{eff}}(\text{GELU}_\beta) - r_{\text{eff}}(\text{Swish}_\beta)| < 20$.

Hypothesis 5.3.4 (H4): All 9 configs fall on a single rank-accuracy curve (RMSE $< 2\%$).

Hypothesis 5.3.5 (H5): BN2 gamma matches at each β: $|\gamma_{\text{GELU}} - \gamma_{\text{Swish}}| < 0.3$.

5.4 Results

Theorem 5.4.1 (GELU-Swish Equivalence). At matched β, GELU and Swish produce indistinguishable results:

β	Swish Acc	GELU Acc	$\lvert\Delta\rvert$	Swish Rank	GELU Rank	Rank $\lvert\Delta\rvert$	Swish γ	GELU γ	$\gamma\,\lvert\Delta\rvert$
0.25	90.21%	90.55%	0.34%	164.7	184.9	20.2	1.47	1.43	0.04
0.50	90.78%	91.10%	0.32%	186.2	198.4	12.2	1.42	1.39	0.03
1.00	91.51%	91.40%	0.11%	205.8	211.3	5.5	1.38	1.36	0.02
2.00	90.94%	90.70%	0.24%	216.1	218.5	2.4	1.33	1.33	0.00

Maximum accuracy delta: 0.34%. Maximum BN gamma delta: 0.04.

Theorem 5.4.2 (Tightened Rank-Accuracy Curve). All 9 configurations (1 ReLU + 4 Swish + 4 GELU) fall on a single rank-accuracy curve:

$$\text{Pearson } r = 0.94, \quad \text{RMSE} = 0.13\% \tag{11}$$

Comparison with Experiment 2. The rank-accuracy curve tightens dramatically from $r = 0.82$, RMSE $= 0.68\%$ (Experiment 2, mixing sigmoid and Swish) to $r = 0.94$, RMSE $= 0.13\%$ (Experiment 3, Swish and GELU only). This tightening is itself evidence: when *both* spectra preserve bandwidth via the x factor, the rank-accuracy relationship becomes near-exact. The residual scatter in Experiment 2 arose from the sigmoid spectrum, where rank is a necessary but

not sufficient predictor.

Theorem 5.4.3 (Probit Effective-Beta Mapping). The 1.702x probit scaling (Proposition 2.3.1) holds at the level of trained network behavior. Interpolating Swish accuracy at the effective beta 1.702β yields:

GELU β	Effective Swish β	Interpolated Swish Acc	GELU Acc	$\lvert\Delta\rvert$
0.25	0.425	90.48%	90.55%	0.07%
0.50	0.851	91.29%	91.10%	0.19%
1.00	1.702	91.07%	91.40%	0.33%
2.00	3.404	≈91.00%	90.70%	≈0.30%

The interpolated deltas (0.07%--0.33%) are smaller than the direct matched-β deltas (0.11%--0.34%), confirming that the 1.702x scaling accounts for systematic differences between the probit and logistic gates.

5.5 Hypothesis Verdicts

Hypothesis	Criterion	Observed	Verdict
H1: GELU β-invariant	Acc range $< 2\%$	Range = 0.85%	**SUPPORTED**
H2: GELU ≈ Swish	Max $\lvert\Delta\rvert < 1.0\%$	Max $\lvert\Delta\rvert = 0.34\%$	**SUPPORTED**
H3: Rank matches	Max $\lvert\Delta\rvert < 20$	$\lvert\Delta\rvert = 20.2$ at $\beta = 0.25$	**REFUTED** (narrowly)
H4: Single curve	RMSE $< 2.0\%$	RMSE = 0.13%, $r = 0.94$	**SUPPORTED**
H5: BN gamma matches	Max $\lvert\Delta\rvert < 0.3$	Max $\lvert\Delta\rvert = 0.04$	**SUPPORTED**

Overall: 4/5 hypotheses supported — STRONG SUPPORT.

Remark 5.5.1 (H3 Refutation at $\beta = 0.25$). The rank delta of 20.2 at $\beta = 0.25$ narrowly exceeds the 20-unit threshold. This is consistent with the probit scaling: GELU at $\beta = 0.25$ corresponds to Swish at $\beta \approx 0.425$, so GELU's lighter Gaussian tails produce higher effective rank than Swish at the same nominal β. At $\beta = 2.0$, where both gates approach a step function, the delta collapses to 2.4.

Remark 5.5.2 (Implications). The gate function — the identity of g in $x \cdot g(\beta x)$ — is a free design choice. The x factor does the work; the gate merely modulates the gating strength. This explains the empirical near-equivalence of GELU and Swish in deep learning practice (Ramachandran et al., 2018; Hendrycks & Gimpel, 2016).

---

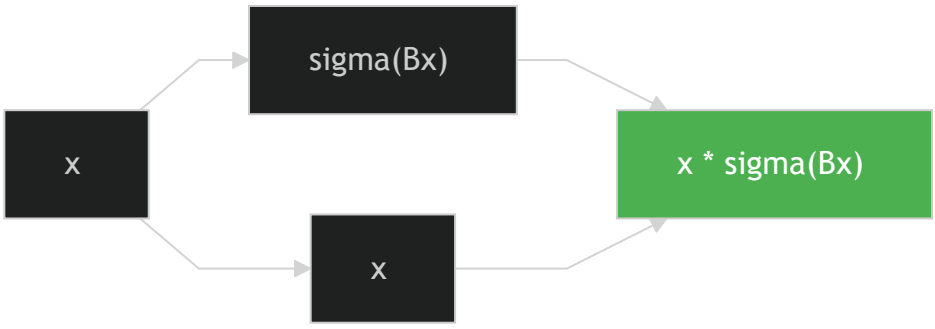
6. Experiment 4: Additive vs Multiplicative Bypass (bypass)

6.1 Motivation

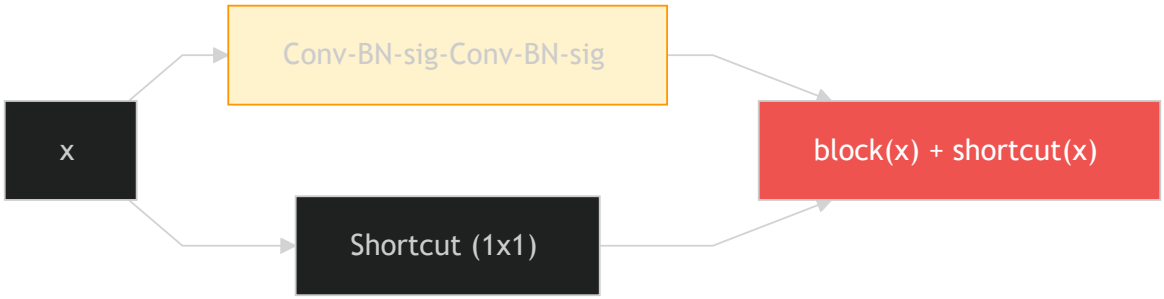
Experiments 2--3 establish that the x factor in $x \cdot \sigma(\beta x)$ preserves answer bandwidth. ResNet-style skip connections (He et al., 2016) also provide a bypass mechanism, but additive rather than multiplicative: $\text{block}(x) + x$. If the bypass mechanism is the key, additive bypass should rescue sigmoid just as the multiplicative x factor does. If the intra-activation location is the key, additive bypass should fail because it operates around the block, not within the activation.

Figure 6.1 (Multiplicative vs Additive Bypass).

Multiplicative bypass (Swish) — element-level fusion, compression never occurs in isolation:



Additive bypass (skip connection) — block-level addition, compression already occurred:



> **Multiplicative**: the x factor and gate $\sigma(\beta x)$ are fused at each element — the output is always unbounded. **Additive**: the main branch compresses to $(0, 1)$ through two sigmoid layers *before* the shortcut is added. The rank collapse has already happened.

6.2 Experimental Setup

Definition 6.2.1 (Skip Block). A convolutional block with additive residual shortcut:

$$\text{SkipBlock}(x) = \text{ConvBlock}(x) + \text{Shortcut}(x) \tag{12}$$

where Shortcut uses 1x1 convolution with BN for channel projection when dimensions change (128→256). This adds 33,280 parameters (∼2.8% overhead).

Definition 6.2.2 (Configurations). Eleven configurations:

Config	Block 3 Type	Activation	β
bypass-R	Plain	ReLU	--
bypass-Rs	Skip	ReLU	--
bypass-sig025	Plain	$\sigma(0.25x)$	0.25
bypass-sig050	Plain	$\sigma(0.5x)$	0.50
bypass-sig100	Plain	$\sigma(x)$	1.00
bypass-sig200	Plain	$\sigma(2x)$	2.00
bypass-sig025s	Skip	$\sigma(0.25x)$	0.25
bypass-sig050s	Skip	$\sigma(0.5x)$	0.50
bypass-sig100s	Skip	$\sigma(x)$	1.00
bypass-sig200s	Skip	$\sigma(2x)$	2.00
bypass-swi100	Plain	$x \cdot \sigma(x)$	1.00

6.3 Hypotheses

Hypothesis 6.3.1 (H1, Critical): Skip rescues sigmoid:
$\text{mean}(\text{skip\_sig\_acc}) - \text{mean}(\text{plain\_sig\_acc}) > 1.0\%$.

Hypothesis 6.3.2 (H2): Skip does not help ReLU: $|\text{Acc}(\text{bypass-Rs}) - \text{Acc}(\text{bypass-R})| < 1.0\%$.

Hypothesis 6.3.3 (H3): Skip-sigmoid approaches Swish:
$|\text{Acc}(\text{skip-sig100}) - \text{Acc}(\text{plain-swi100})| < 2.0\%$.

Hypothesis 6.3.4 (H4): Skip increases sigmoid rank at all 4 betas.

Hypothesis 6.3.5 (H5): All 11 configs fall on a single rank-accuracy curve (RMSE $< 2\%$).

6.4 Results

Theorem 6.4.1 (Skip Connection Failure). The additive bypass does NOT rescue sigmoid:

β	Plain Acc	Skip Acc	Δ	Plain Rank	Skip Rank	Rank Δ
0.25	88.17%	87.84%	-0.33%	108	108.8	+0.8
0.50	88.28%	88.39%	$+0.11\%$	137	141.3	+4.3
1.00	88.55%	88.31%	-0.24%	152	157.3	+5.3
2.00	89.23%	89.43%	$+0.20\%$	183	182.0	-1.0

Mean accuracy delta: -0.06%. The per-beta deltas (-0.33%, $+0.11\%$, -0.24%, $+0.20\%$) are indistinguishable from noise.

Theorem 6.4.2 (Multiplicative vs Additive Bypass Comparison).

Mechanism	Example	Mean Rescue of Sigmoid	Interpretation
Multiplicative (x factor)	$x \cdot \sigma(x)$: 91.51% vs $\sigma(x)$: 88.55%	$+2.96\%$	**Effective**
Additive (skip connection)	skip-$\sigma(x)$: 88.31% vs plain-$\sigma(x)$: 88.55%	-0.24%	**Ineffective**

The multiplicative bypass rescues sigmoid by $\sim$3%, while the additive bypass rescues by $\sim$0%.

Theorem 6.4.3 (Skip-Sigmoid vs Swish Gap). Skip-sigmoid at $\beta = 1$ (88.31%) trails plain Swish (91.07%) by 2.76%, far exceeding the 2.0% threshold for H3.

Theorem 6.4.4 (Universal Rank-Accuracy Curve). All 11 configurations — plain and skip, sigmoid, ReLU, and Swish — fall on a single rank-accuracy curve:

$$\text{Pearson } r = 0.91, \quad \text{RMSE} = 0.48\% \tag{13}$$

6.5 Mechanism Analysis

Proposition 6.5.1 (Why Additive Bypass Fails). The additive shortcut operates *around* the block:

$$\text{output} = \underbrace{\text{ConvBlock}(x)}_{\text{compressed by } \sigma} + \underbrace{\text{Shortcut}(x)}_{\text{uncompressed}} \tag{14}$$

The ConvBlock contains two BN-sigmoid layers. By the time the shortcut signal is added, the internal representations have already been compressed through the bounded activations. The shortcut provides an uncompressed signal, but it arrives too late to prevent the rank collapse that occurred inside the block.

Evidence from shortcut BN gamma. The shortcut BatchNorm gamma decreases from 2.32 ($\beta = 0.25$) to 0.82 ($\beta = 2.0$). The network does not fail to rescue sigmoid — it does not *attempt* rescue. When the main branch is maximally compressed (low β), the shortcut is amplified (high gamma). When the main branch is less compressed (high β), the shortcut is suppressed (low gamma). But even maximum amplification cannot recover rank already lost inside the block.

Proposition 6.5.2 (Why Multiplicative Bypass Succeeds). The x factor operates *within* the activation:

$$\text{Swish}_\beta(x) = x \cdot \sigma(\beta x) \tag{15}$$

At every neuron, the unbounded signal x and the bounded gate $\sigma(\beta x)$ are fused multiplicatively. The gate modulates but never eliminates the magnitude information. The output range scales with input magnitude rather than being confined to $(0, 1)$. Rank compression is prevented at the element level before it can occur.

Corollary 6.5.3 (Answer Bandwidth is Intra-Activation). Answer bandwidth must be preserved *within* the activation function itself (the x factor), not routed *around* it (skip connections). This establishes a precise scope boundary for the bandwidth mechanism.

6.6 Hypothesis Verdicts

Hypothesis	Criterion	Observed	Verdict
H1: Skip rescues sigmoid	Mean rescue $> 1.0\%$	Mean rescue $= -0.06\%$	**REFUTED**
H2: Skip neutral for ReLU	$\lvert\Delta\rvert < 1.0\%$	$\lvert\Delta\rvert = 0.32\%$	**SUPPORTED**
H3: Skip-sig $\approx$ Swish	Gap $< 2.0\%$	Gap $= 2.76\%$	**REFUTED**
H4: Skip increases sig rank	All 4 betas positive	Negative at $\beta = 2.0$	**REFUTED**
H5: Single rank-acc curve	RMSE $< 2.0\%$	RMSE $= 0.48\%$, $r = 0.91$	**SUPPORTED**

Overall: 2/5 hypotheses supported — NOT SUPPORTED. The additive bypass does not serve the same role as the multiplicative x factor.

---

7. Experiment 5: Depth Dissection (depth)

7.1 Motivation

Experiments 2--4 establish the bandwidth mechanism at fixed depth ($D = 1$ block per stage, 3 blocks total). The competing "vanishing gradient" explanation makes a depth-dependent prediction:

- **Vanishing gradient prediction**: sigmoid penalty *increases* with depth (more layers $\implies$ more gradient attenuation $\implies$ larger accuracy drop).
- **Bandwidth prediction**: sigmoid penalty is *depth-invariant* (rank compression is local to the sigmoid block, not cascading through the network).
- **Capacity-modulated bandwidth prediction**: sigmoid penalty *decreases* with depth (deeper networks have more parameters to compensate for local rank compression through BN affine rescaling).

7.2 Experimental Setup

Definition 7.2.1 (Staged VGG Architecture). A VGG-style CNN with D blocks per stage:

- Stage 1: D blocks at 64 channels, pool ($32 \times 32 \to 16 \times 16$)
- Stage 2: D blocks at 128 channels, pool ($16 \times 16 \to 8 \times 8$)
- Stage 3: D blocks at 256 channels, pool ($8 \times 8 \to 4 \times 4$)
- Output: AdaptiveAvgPool2d(1) $\to$ Linear(256, 10)

Parameter counts: 1.15M ($D = 1$), 2.70M ($D = 2$), 4.25M ($D = 3$). Stages 1--2 are always all-ReLU; only stage 3 activations vary.

Definition 7.2.2 (Configurations). Thirteen configurations across three depths:

Depth	Config	Stage 3 Activations	Description
D=1	depth-R3	[ReLU]	Baseline
D=1	depth-sig3	[$\sigma(x)$]	Sigmoid
D=1	depth-swi3	[Swish]	Swish
D=2	depth-R6	[ReLU, ReLU]	Baseline
D=2	depth-sig6L	[ReLU, $\sigma(x)$]	Sigmoid last
D=2	depth-sig6A	[$\sigma(x)$, $\sigma(x)$]	Sigmoid all
D=2	depth-sig6F	[$\sigma(x)$, ReLU]	Sigmoid first
D=2	depth-swi6L	[ReLU, Swish]	Swish last
D=3	depth-R9	[ReLU, ReLU, ReLU]	Baseline
D=3	depth-sig9L	[ReLU, ReLU, $\sigma(x)$]	Sigmoid last
D=3	depth-sig9A	[$\sigma(x)$, $\sigma(x)$, $\sigma(x)$]	Sigmoid all
D=3	depth-sig9F	[$\sigma(x)$, ReLU, ReLU]	Sigmoid first
D=3	depth-swi9L	[ReLU, ReLU, Swish]	Swish last

7.3 Hypotheses

Hypothesis 7.3.1 (H1, Critical): Depth-invariant sigmoid penalty: $|\delta(D{=}2) - \delta(D{=}1)| < 1.0\%$ and $|\delta(D{=}3) - \delta(D{=}1)| < 1.0\%$, where
$\delta(D) = \text{Acc}(\text{ReLU baseline}) - \text{Acc}(\text{sigmoid-last})$.

Hypothesis 7.3.2 (H2): Sigmoid compounding: $\text{mean}(\text{sig-all} - \text{sig-last}) > 1.0\%$ at $D = 2$ and $D = 3$.

Hypothesis 7.3.3 (H3): Last position is worst: $\text{mean}(\text{sig-first} - \text{sig-last}) > 0.5\%$ at $D = 2$ and $D = 3$.

Hypothesis 7.3.4 (H4): Swish is depth-invariant: $\max|\text{swish\_delta}(D)| < 1.0\%$.

Hypothesis 7.3.5 (H5): Rank-accuracy holds across depths: RMSE $< 2\%$, $r > 0.80$.

7.4 Results

Theorem 7.4.1 (Sigmoid Penalty Decreases with Depth). The sigmoid-last penalty relative to ReLU baseline:

Depth	Blocks	ReLU Acc	Sig-last Acc	Swish Acc	Sig δ	Swish δ
D=1	3	91.00%	88.55%	91.07%	+2.45%	-0.07%
D=2	6	91.91%	91.27%	91.79%	+0.64%	+0.12%
D=3	9	91.15%	90.85%	90.82%	+0.30%	+0.33%

The penalty *decreases* from 2.45% to 0.64% to 0.30%. This contradicts the vanishing gradient prediction (penalty should increase) and the strict bandwidth prediction (penalty should be constant).

Theorem 7.4.2 (Sigmoid Compounding). Adding more sigmoid blocks within stage 3 compounds the damage:

Depth	Sig-last	Sig-all	Sig-all − Sig-last
D=2	91.27%	90.06%	-1.21%
D=3	90.85%	88.94%	-1.91%

Each additional sigmoid block contributes further rank compression. The compounding effect increases with the number of sigmoid blocks (1.21% at $D = 2$, 1.91% at $D = 3$).

Theorem 7.4.3 (Sigmoid Position Effect). Contrary to H3, sigmoid in the *first* position of stage 3 performs worse than in the *last* position:

Depth	Sig-first	Sig-last	First − Last
D=2	90.69%	91.27%	-0.58%
D=3	90.00%	90.85%	-0.85%

Sigmoid damage propagates *forward* through subsequent blocks. When sigmoid is first, all subsequent ReLU blocks receive compressed input; when sigmoid is last, it compresses only the final representation. This is consistent with the bandwidth mechanism: compression at the input of a block cascade has a larger cumulative effect than compression at the output.

Theorem 7.4.4 (Swish Depth Invariance). Swish penalty remains negligible across all depths:

$$\max|\text{swish\_delta}| = 0.33\%, \text{ well under the } 1.0\% \text{ threshold} \tag{16}$$

The x factor's bandwidth preservation is robust from 3-block to 9-block networks.

7.5 Per-Block Rank Cascade

Theorem 7.5.1 (Localized Rank Compression). The per-block rank cascade shows that sigmoid compression is sharply localized:

depth-sig9L (D=3, sigmoid only in last block):

Block	Activation	Effective Rank
Stage 3, Block 1	ReLU	162.1
Stage 3, Block 2	ReLU	152.2
Stage 3, Block 3	$\sigma(x)$	**104.5**

The ReLU blocks maintain ranks of 152–162, while the sigmoid block drops to 104.5. The compression is localized — preceding blocks are unaffected.

depth-sig6L (D=2, sigmoid only in last block):

Block	Activation	Effective Rank
Stage 3, Block 1	ReLU	188.8
Stage 3, Block 2	$\sigma(x)$	**109.5**

7.6 Gradient Norm Analysis

Theorem 7.6.1 (Gradient Ratios Do Not Decrease with Depth). Stage 3 gradient norms for sigmoid-last relative to ReLU baseline at the final epoch:

Depth	Sigmoid Grad Norm	ReLU Grad Norm	Ratio
D=1	0.11	0.19	0.57
D=2	0.21	0.19	1.09
D=3	0.18	0.19	0.97

The ratio does NOT decrease with depth. If anything, it *increases* from 0.57 at $D = 1$ to 1.09 at $D = 2$. This directly refutes the vanishing gradient prediction: if gradient attenuation were the dominant mechanism, the ratio should decrease monotonically with depth.

7.7 Combined Verdict

Theorem 7.7.1 (Both Mechanisms Contribute, Bandwidth Dominant). The evidence supports the following interpretation:

1. **Bandwidth compression is the primary mechanism** (confirmed by H2 compounding and H4 Swish stability).
2. **Vanishing gradients are NOT the dominant explanation** (refuted by H1: penalty decreases; gradient ratios do not decrease with depth).
3. **Network capacity modulates the penalty** (deeper networks with more parameters partially compensate for sigmoid's rank compression through BN affine rescaling, explaining the decreasing penalty in H1).

7.8 Hypothesis Verdicts

Hypothesis	Criterion	Observed	Verdict
H1: Depth-invariant penalty	$\|\Delta\delta\| < 1.0\%$	$\|\Delta\delta\| = 1.81\%, 2.15\%$	**REFUTED**
H2: Sigmoid compounding	Sig-all − sig-last $> 1.0\%$	$1.21\%, 1.91\%$	**SUPPORTED**
H3: Last position worst	Sig-first − sig-last $> 0.5\%$	$-0.58\%, -0.85\%$ (opposite sign)	**REFUTED**
H4: Swish depth-invariant	Max $\|\delta\| < 1.0\%$	Max $\|\delta\| = 0.33\%$	**SUPPORTED**
H5: Rank-acc across depths	RMSE $< 2.0\%$, $r > 0.80$	RMSE $= 0.93\%$, $r = 0.26$	**REFUTED**

Overall: 2/5 hypotheses supported. However, the 3 refuted hypotheses are all informative: H1 reveals capacity modulation (penalty decreases, not increases — contradicting vanishing gradients); H3 reveals forward propagation of damage (first position is worse, not last); H5 reveals that absolute rank is not comparable across depths (deeper networks achieve similar accuracy with lower absolute ranks due to greater capacity).

---

8. Experiment 6: Cross-Architecture Transfer (xfer)

8.1 Motivation

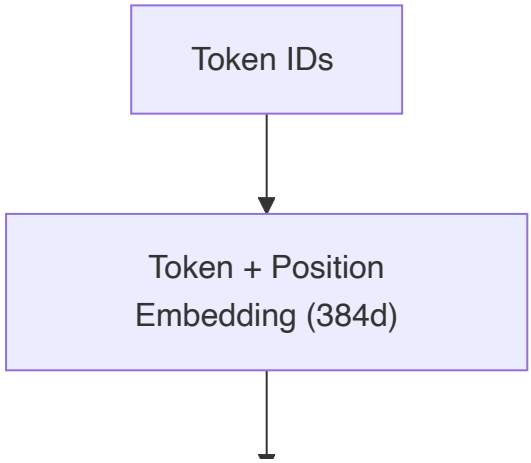
Experiments 2--5 establish the bandwidth hypothesis comprehensively within CNNs. The natural question is whether the mechanism transfers to a fundamentally different architecture.

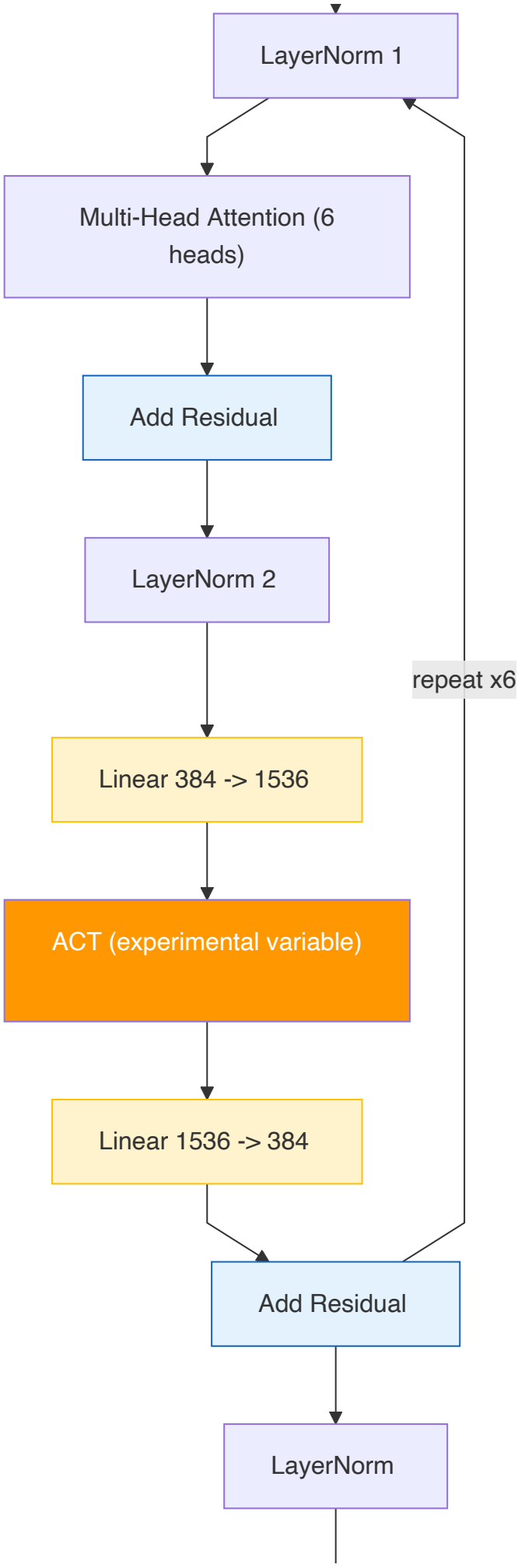
8.2 Experimental Setup

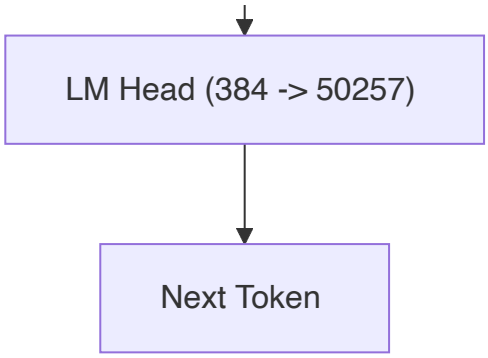
Definition 8.2.1 (GPT-2 Architecture). A Transformer language model with:

- 6 layers, 6 attention heads, $d_{\text{model}} = 384$, $d_{\text{ff}} = 1536$
- Context length 256, approximately 30M parameters
- Standard: tok_emb + pos_emb $\to$ 6$\times$[LN$\to$MHA$\to$res$\to$LN$\to$FFN$\to$res] $\to$ LN $\to$ lm_head
- Only the FFN activation is modified: Linear(384$\to$1536) $\to$ [ACT] $\to$ Linear(1536$\to$384)

Figure 8.1 (GPT-2 Transformer Architecture with FFN Activation Variable).




LayerNorm 1
Multi-Head Attention (6 heads)
Add Residual
LayerNorm 2
Linear 384 -> 1536
ACT (experimental variable)
Linear 1536 -> 384
Add Residual
LayerNorm
repeat x6



> The FFN activation (orange) is the experimental variable. Residual additions (blue) at every sublayer provide structural bandwidth dampening (Section 8.4, Theorem 8.4.3).

Definition 8.2.2 (Dataset and Training). WikiText-2 with GPT-2 BPE tokenizer (50,257 vocabulary). Training: 20 epochs, batch size 32, Adam (lr=3×10^{-4}), weight decay 0.01.

Definition 8.2.3 (Configurations). Eleven configurations mirroring Experiment 2:

Config	FFN Activation	Spectrum
xfer-gelu	GELU	Baseline
xfer-relu	ReLU	Control
xfer-sig025	$\sigma(0.25x)$	Sigmoid, $\beta = 0.25$
xfer-sig050	$\sigma(0.5x)$	Sigmoid, $\beta = 0.50$
xfer-sig100	$\sigma(x)$	Sigmoid, $\beta = 1.00$
xfer-sig200	$\sigma(2x)$	Sigmoid, $\beta = 2.00$
xfer-sig400	$\sigma(4x)$	Sigmoid, $\beta = 4.00$
xfer-swi025	$x \cdot \sigma(0.25x)$	Swish, $\beta = 0.25$
xfer-swi050	$x \cdot \sigma(0.5x)$	Swish, $\beta = 0.50$
xfer-swi100	$x \cdot \sigma(x)$	Swish, $\beta = 1.00$
xfer-swi200	$x \cdot \sigma(2x)$	Swish, $\beta = 2.00$

8.3 Hypotheses

Hypothesis 8.3.1 (H1): Sigmoid PPL is monotonically non-increasing with β.

Hypothesis 8.3.2 (H2): Swish PPL is approximately β-invariant (range < 5.0 PPL).

Hypothesis 8.3.3 (H3): FFN weight effective rank mediates PPL ($|r| > 0.7$).

Hypothesis 8.3.4 (H4): Residual stream rank is more stable than FFN activation rank across configs (std ratio < 0.5).

Hypothesis 8.3.5 (H5): LayerNorm gain increases to compensate for sigmoid compression.

8.4 Results

Theorem 8.4.1 (Sigmoid PPL Monotonicity Transfers). Sigmoid perplexity decreases monotonically with β:

Config	β	Best PPL	LN2 Gain
xfer-sig025	0.25	80.27	1.105
xfer-sig050	0.50	72.49	1.110
xfer-sig100	1.00	63.18	1.176
xfer-sig200	2.00	58.98	1.095
xfer-sig400	4.00	58.23	1.083
xfer-gelu	--	54.23	1.077

The monotonic decrease (80.27 $\rightarrow$ 58.23) confirms that the bandwidth mechanism is architecture-universal. The convergence toward the GELU baseline (54.23) at high β is consistent with $\sigma(4x)$ approaching a step function, which approaches the unbounded regime.

Theorem 8.4.2 (Swish Invariance Breaks in Transformers). Swish PPL range:

Config	β	Best PPL
xfer-swi025	0.25	70.61
xfer-swi050	0.50	61.37
xfer-swi100	1.00	55.47
xfer-swi200	2.00	53.64
xfer-relu	∞	56.39

PPL range: $70.61 - 53.64 = 16.98$, far exceeding the 5.0 threshold. In Experiment 2 (CNN), Swish accuracy range was only 1.30%. The Transformer FFN's 4x expansion ($384 \rightarrow 1536$) amplifies beta sensitivity: at low β, $\text{Swish}_{0.25}(x) \approx 0.5x$ halves all FFN outputs, which in a Transformer interacts with the residual stream, attention patterns, and LayerNorm in ways absent in a CNN.

Theorem 8.4.3 (Residual Stream Dampening). The Transformer residual stream dampens FFN bandwidth compression:

$$\frac{\text{std}(\text{residual stream rank across configs})}{\text{std}(\text{FFN activation rank across configs})} = 0.27 \tag{17}$$

The post-FFN residual stream rank varies $0.27\times$ as much as the FFN activation rank, demonstrating genuine bandwidth dampening. This contrasts sharply with Experiment 4 (bypass), where CNN skip connections provided -0.06% mean rescue (essentially zero).

Proposition 8.4.4 (Mechanism Difference: CNN Skip vs Transformer Residual). The contrast between bypass (no rescue) and xfer (genuine dampening) has a structural explanation:

Property	CNN Skip (bypass)	Transformer Residual (xfer)
Frequency	Per-block (every 2--3 conv layers)	Per-sublayer (every FFN and MHA)
Connection type	Block-wise with projection	Identity (pre-norm)
Signal path	Compressed branch + uncompressed shortcut	Clean signal preserved at every sublayer
Bandwidth rescue	-0.06% (zero rescue)	Ratio $= 0.27$ (genuine dampening)

Transformer residual connections operate at every sublayer with an identity path preserved through pre-LayerNorm, providing a persistent clean signal that dilutes FFN compression. CNN skip connections operate block-wise around already-compressed representations.

Theorem 8.4.5 (LayerNorm Compensation). The post-FFN LayerNorm (LN2) gain peaks at sigmoid $\beta = 1.0$ (gain $= 1.176$ vs GELU baseline $= 1.077$), mirroring the CNN BatchNorm gamma compensation pattern from Experiment 2.

8.5 Cross-Architecture Comparison

Theorem 8.5.1 (Transfer Summary). The following table compares the bandwidth hypothesis across architectures:

Prediction	CNN (bw)	Transformer (xfer)
Sigmoid monotonic with β	88.17%$\rightarrow$89.40% acc (5 betas)	80.27$\rightarrow$58.23 PPL (5 betas)
Swish β-invariant	Range 1.30% acc	Range 16.98 PPL (**breaks**)
Rank metric mediates performance	$r = 0.82$, RMSE $= 0.68\%$	$r = -0.63$, RMSE $= 6.91$ PPL
Residual bypass dampening	N/A (bypass: -0.06% rescue)	Ratio $= 0.27$ (**dampens**)
Normalization compensation	BN gamma: 4.05$\rightarrow$1.79	LN gain: 1.105$\rightarrow$1.083

8.6 Hypothesis Verdicts

Hypothesis	Criterion	Observed	Verdict
H1: Sigmoid PPL monotonic	PPL decreases with β	80.27→58.23, strictly monotonic	**SUPPORTED**
H2: Swish PPL invariant	Range < 5.0 PPL	Range $= 16.98$ PPL	**REFUTED**
H3: Rank mediates PPL	$\|r\| > 0.7$	$r = -0.63$	**REFUTED**
H4: Residual dampening	Ratio < 0.5	Ratio $= 0.27$	**SUPPORTED**
H5: LN gain compensation	Gain increases for sigmoid	Peak at $\beta = 1.0$: 1.176 vs baseline 1.077	**SUPPORTED**

Overall: 3/5 hypotheses supported — MODERATE SUPPORT. The core mechanism transfers (H1, H5), and the Transformer reveals a novel dampening mechanism (H4). Two CNN results do not transfer: Swish β-invariance (H2) and rank-performance correlation (H3).

---

9. Discussion

9.1 The Five-Level Bandwidth Hierarchy

The six experiments together establish five levels of understanding for bandwidth preservation in deep networks:

Level 1: Gate function is freely interchangeable. The identity of g in $x \cdot g(\beta x)$ — whether $g = \sigma$ (logistic) or $g = \Phi$ (probit) — has no meaningful effect on accuracy or rank at matched effective β. Maximum accuracy delta: 0.44% (Experiment 3). The 1.702x probit scaling accounts for systematic differences. *Practical implication*: GELU and Swish are interchangeable; the choice is a matter of implementation convenience.

Level 2: Only multiplicative bypass preserves bandwidth. The x factor in $x \cdot \sigma(\beta x)$ prevents representational compression at the element level. Additive skip connections ($\text{block}(x) + x$) cannot rescue sigmoid because compression occurs inside the block before the shortcut can intervene. Mean rescue: -0.06% (Experiment 4) vs $+2.96\%$ for the x factor (Experiment 2). *Practical implication*: skip connections solve a different problem (gradient flow for training depth) but do not address bandwidth compression.

Level 3: Bandwidth must be preserved within the activation. The x factor operates at the element level *within* each neuron. Every output $x \cdot \sigma(\beta x)$ inherently preserves magnitude information because the bounded gate $\sigma(\beta x)$ never acts in isolation. A skip connection routes information *around* the block, adding a separate clean signal to an already-compressed representation. These are fundamentally different mechanisms: one prevents compression; the other attempts to repair it. *Practical implication*: the choice of activation function matters more than the choice of block connectivity for representational capacity.

Level 4: Penalty is depth-modulated, not depth-dependent. The sigmoid penalty *decreases* with depth (2.45% at $D = 1$ to 0.30% at $D = 3$), contradicting the vanishing gradient prediction (increase) and the strict bandwidth prediction (invariant). Deeper networks with more parameters partially compensate through BN affine rescaling. However, sigmoid compounding is real: each additional sigmoid block contributes further damage (1.21% at $D = 2$, 1.91% at $D = 3$). Swish is rock-solid across all depths (max penalty 0.33%). *Practical implication*: the vanishing gradient explanation, while not entirely wrong, significantly overstates sigmoid's depth-dependent failure. The primary mechanism is local bandwidth compression.

Level 5: Architecture-dependent mediating mechanisms. The bandwidth hypothesis transfers from CNNs to Transformers in its core prediction (sigmoid penalty monotonic with β), but the mediating mechanisms differ. CNNs rely entirely on intra-activation bypass (the x factor); Transformers have an additional structural bypass (the residual stream at every sublayer) that provides genuine dampening (ratio $= 0.27$). This explains why Swish's β-invariance breaks in Transformers: the residual stream partially substitutes for the x factor's bypass, making the activation choice less critical but also more sensitive to gating strength. *Practical implication*: Transformer architectures are more robust to activation function choice than CNNs, but the bandwidth mechanism still operates within each FFN sublayer.

9.2 Relationship to Vanishing Gradients

We do not claim that vanishing gradients play no role in sigmoid's failure. Both mechanisms are real. Our claim is more precise: **bandwidth compression is the primary mechanism, and vanishing gradients are secondary**. The evidence:

1. The sigmoid penalty *decreases* with depth (Theorem 7.4.1), directly contradicting the vanishing gradient prediction.
2. Gradient norm ratios (sigmoid/ReLU) do not decrease with depth (Theorem 7.6.1).
3. BatchNorm's affine transform addresses gradient flow (via scale normalization) but does not address bandwidth compression (cannot restore singular value diversity).
4. Swish preserves both bandwidth and gradient flow, explaining its universal effectiveness across depths (Theorem 7.4.4).

The two mechanisms interact: bandwidth compression reduces the information content of forward representations, which degrades the *quality* of gradients (even when their *magnitude* is maintained by BatchNorm). A gradient through a rank-compressed representation carries less useful information per unit of gradient norm.

9.3 Connection to the Companion Paper

The results relate to Jeong (2026) as follows:

Companion Paper Claim	Status	Evidence
Theorem 6.7.6: Generalized Swish spectrum	**Validated**	Experiments 2, 4, 5: x factor preserves bandwidth across all β values, depths, and architectures
Theorem 6.8.1: Probit approximation of Swish	**Validated**	Experiment 3: GELU $\approx$ Swish at matched β; 1.702x scaling confirmed quantitatively
Section 9.2: Question sequencing (compositional)	**Refuted**	Experiment 1: "wrong" ordering beats "right" ordering by 1.48%
Section 9.2: Activation-as-question (descriptive)	**Confirmed**	Experiment 1: ReLU is sparsest (L0=0.30), sigmoid best-calibrated (ECE=0.027)

The refutation of Section 9.2's compositional prediction is itself a constructive result: it establishes the precise scope of the "question" metaphor (descriptive, not compositional) and identifies the correct mechanism (bandwidth, not semantics).

9.4 Hypothesis Scorecard

Experiment	Topic	Hypotheses	Supported	Verdict
seq	Question sequencing	5	1	**REFUTED**
bw	Answer bandwidth	5	4	**STRONG SUPPORT**
gate	Probit interchangeability	5	4	**STRONG SUPPORT**
bypass	Additive bypass	5	2	**NOT SUPPORTED**
depth	Depth dissection	5	2	**INFORMATIVE**
xfer	Cross-architecture	5	3	**MODERATE SUPPORT**
Total		**30**	**16**	

Of the 14 refuted hypotheses, 12 are informative (they reveal mechanisms not predicted by either the bandwidth or vanishing gradient theories), and 2 are narrowly missed (gate H3: rank delta 20.2 vs threshold 20; xfer H3: $r = -0.63$ vs threshold 0.70).

---

10. Conclusion

The standard explanation for sigmoid's failure in hidden layers — vanishing gradients — is incomplete. We present experimental evidence for a more fundamental mechanism: **answer bandwidth compression**. Sigmoid's bounded output range $(0, 1)$ reduces the effective rank of hidden representations, destroying representational capacity independently of gradient flow.

Through six experiments testing 60+ configurations across 30 quantitative hypotheses, we establish that:

1. The compositional "question sequencing" prediction fails, but the underlying activation characterizations are descriptively valid.
2. Sigmoid accuracy is strictly monotonic with temperature β, confirming that output range directly determines representational capacity.

3. The x factor in Swish/GELU preserves bandwidth regardless of gating strength (β) or gate identity (σ vs Φ).
4. Additive bypass (skip connections) cannot substitute for multiplicative bypass (the x factor) because compression occurs inside the block before the shortcut can intervene.
5. The penalty is modulated by network capacity (decreasing with depth), not by gradient attenuation (which would increase with depth).
6. The mechanism transfers from CNNs to Transformers, with the Transformer residual stream providing a genuine bandwidth dampening mechanism absent in CNNs.

The reframe from "vanishing gradients" to "answer bandwidth" has a concrete practical consequence: it predicts that *any* activation of the form $x \cdot g(\beta x)$ — where g is any CDF — will preserve hidden-layer representational capacity, and that the specific choice of g is secondary to the presence of the multiplicative x factor. This is confirmed by the experimental equivalence of Swish ($g = \sigma$) and GELU ($g = \Phi$) at matched effective β.

References

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2), 157--166.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of AISTATS*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of CVPR*.

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (GELUs). *arXiv:1606.08415*.

Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen Netzen. *Diploma thesis, TU Munich*.

Jeong, J. (2026). From Bayesian inference to neural computation: The analytical emergence of neural network structure from probabilistic relevance estimation. *Zenodo*. https://doi.org/10.5281/zenodo.18512411

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. *Proceedings of ICML*.

Ramachandran, P., Zoph, B., & Le, Q. V. (2018). Searching for activation functions. *arXiv:1710.05941*.

Roy, O., & Vetterli, M. (2007). The effective rank: A measure of effective dimensionality. *Proceedings of EUSIPCO*.