

Activation Functions

Comprehensive Demo and Visualization Report

ReLU | LeakyReLU | Sigmoid | Tanh | GELU | SiLU/Swish

Seed: 42 | NumPy from-scratch implementation

Summary of Key Findings

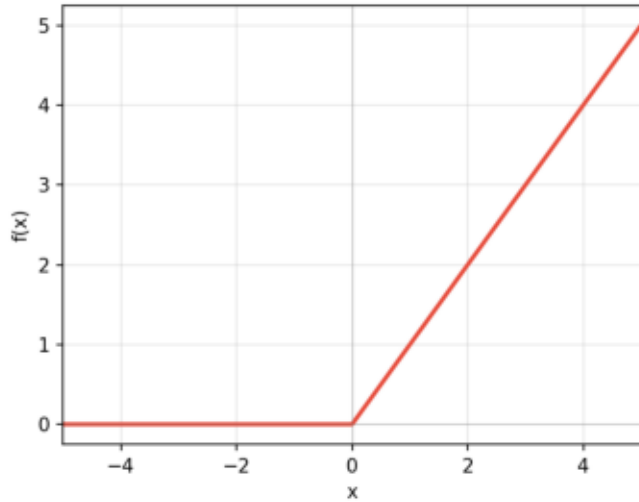
=====

1. Forward Functions: All 6 activations plotted over $[-5, 5]$.
GELU and SiLU show smooth, non-monotonic transitions.
2. Derivatives: Sigmoid and Tanh derivatives peak at 0.25 and 1.0 respectively, vanishing for large $|x|$. ReLU is discontinuous.
3. Dying ReLU: ReLU has 100% zero gradients for negative inputs.
LeakyReLU, GELU, and SiLU maintain non-zero gradients.
4. GELU Approximation: Tanh approximation matches exact (erf) with max absolute error < 0.005 over $[-5, 5]$.
5. Vanishing Gradients: Sigmoid and Tanh saturate at extremes.
Modern activations (ReLU, GELU, SiLU) maintain gradient flow for positive inputs.
6. Gradient Flow: Through 10 chained layers, Sigmoid gradients vanish exponentially. GELU/SiLU maintain better flow.
7. Temperature Scaling: Higher input scaling sharpens activation transitions – Sigmoid becomes step-like, GELU becomes ReLU-like.
8. Numerical Stability: All implementations handle extreme values ($|x|$ up to 1000) without NaN or Inf.

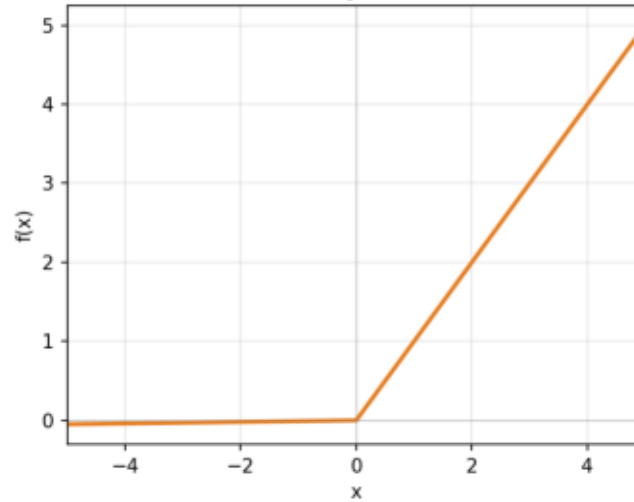
Example 1: Forward Functions (Individual)

Activation Functions — Forward Pass

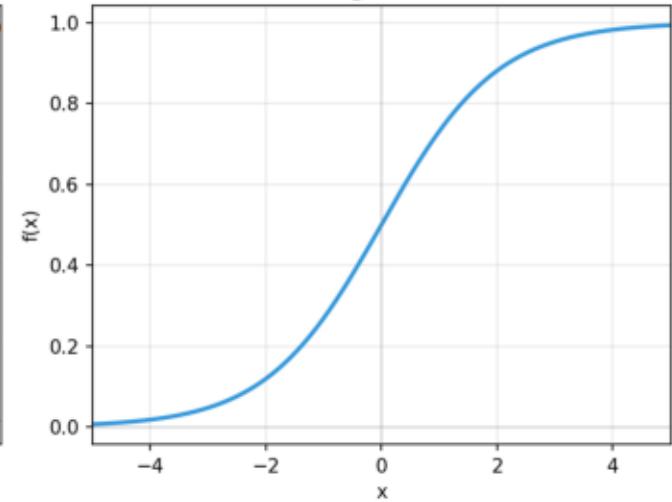
ReLU



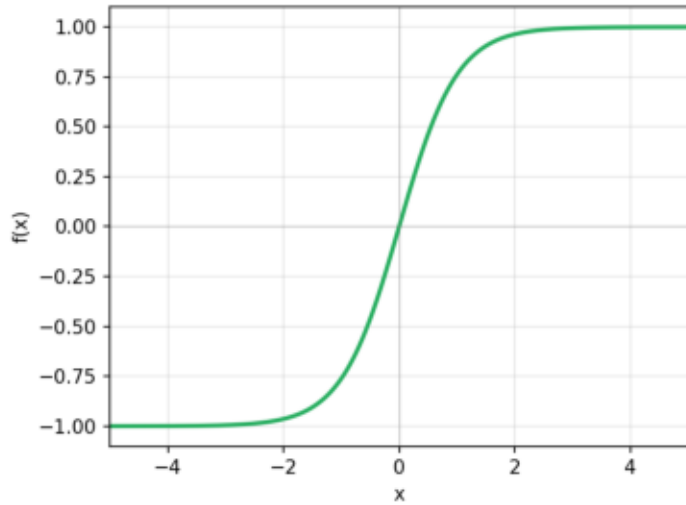
LeakyReLU



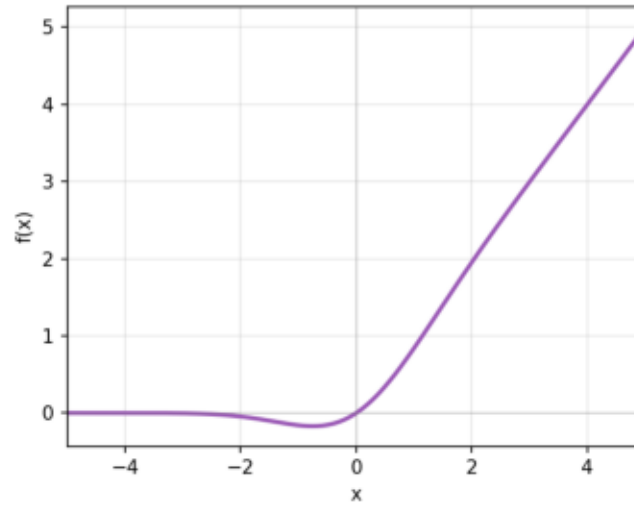
Sigmoid



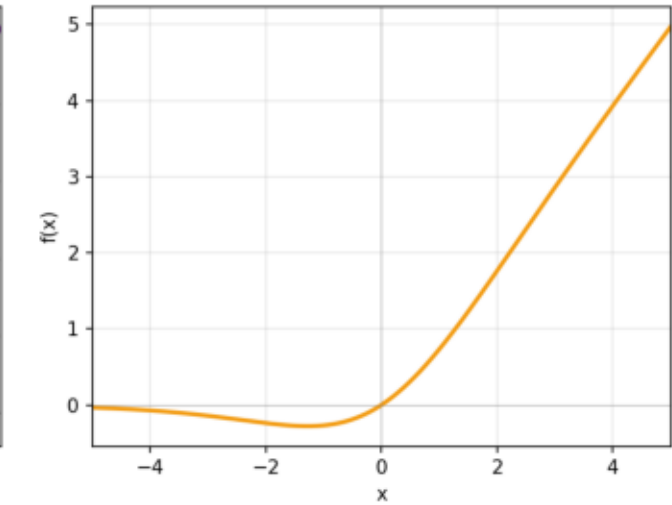
Tanh



GELU

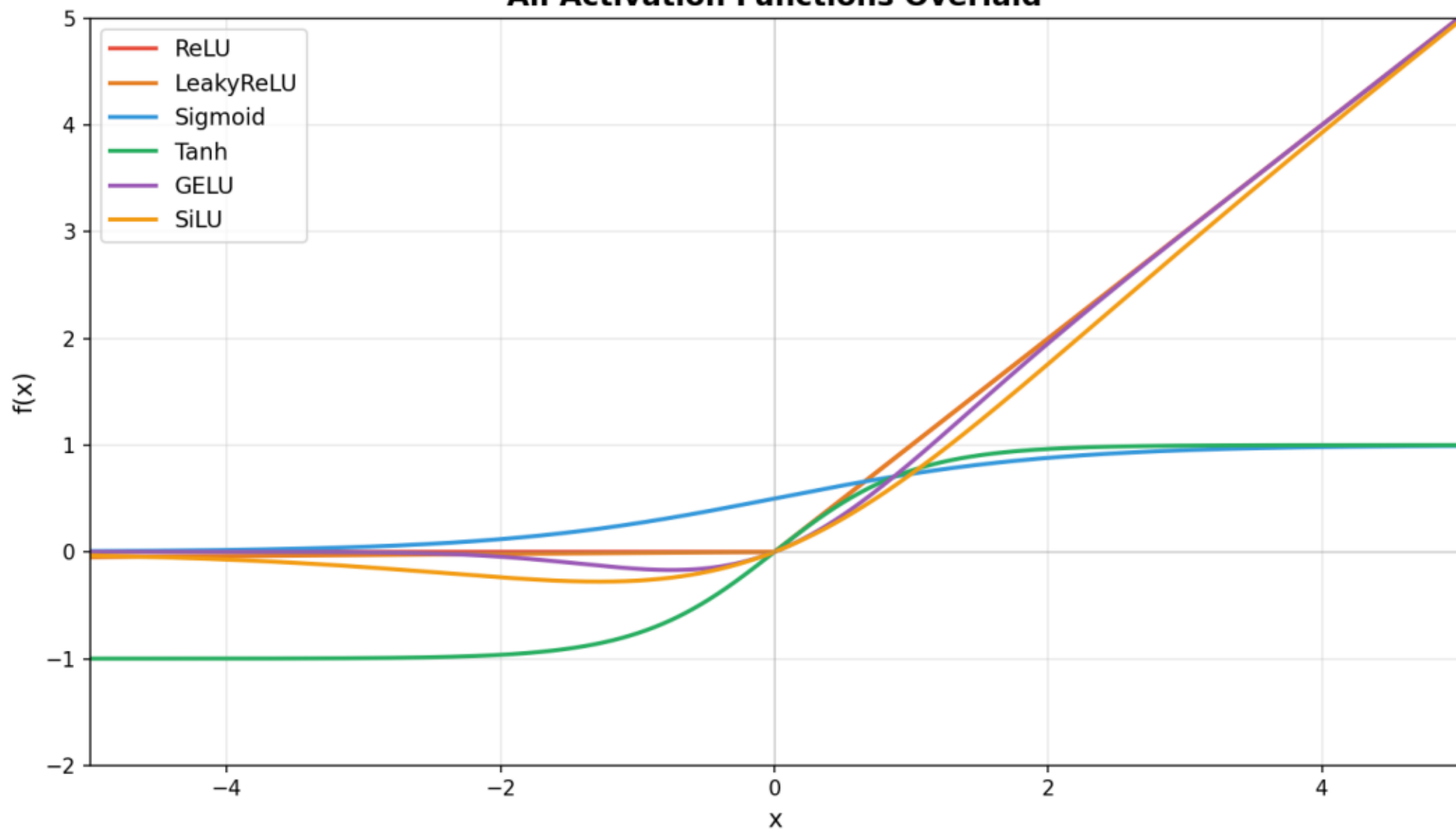


SiLU



Example 1b: Forward Functions (Overlay)

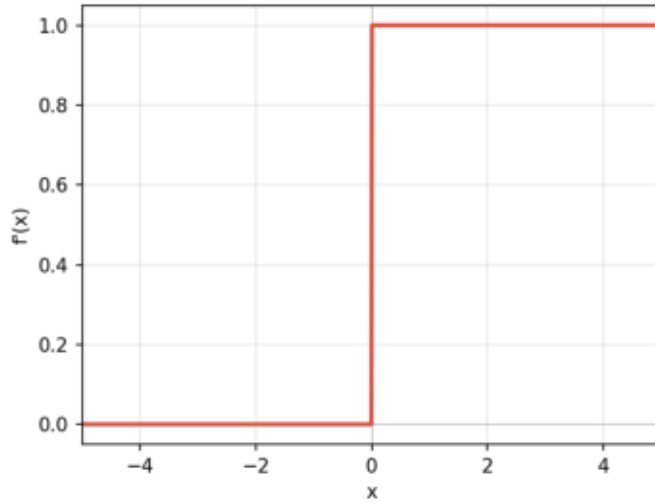
All Activation Functions Overlaid



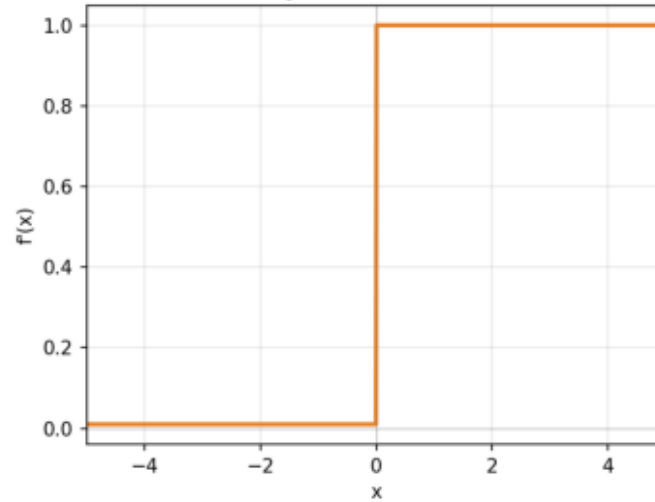
Example 2: Derivatives (Individual)

Activation Derivatives — Backward Pass

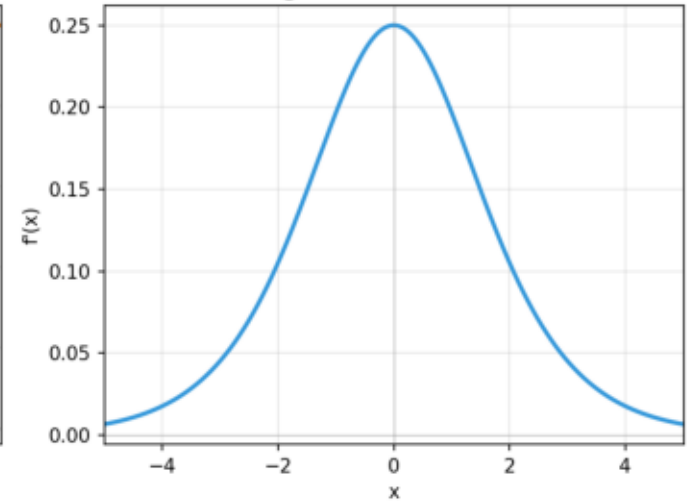
ReLU derivative



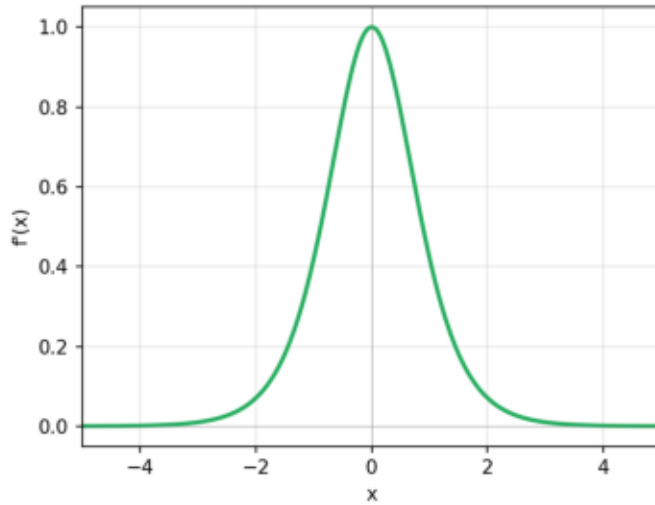
LeakyReLU derivative



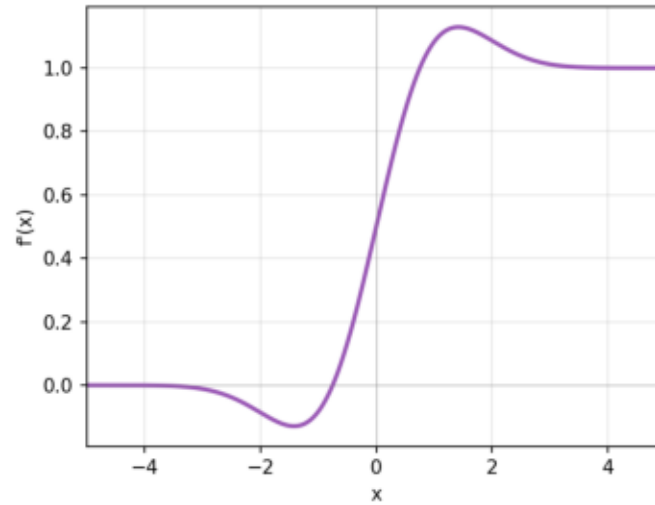
Sigmoid derivative



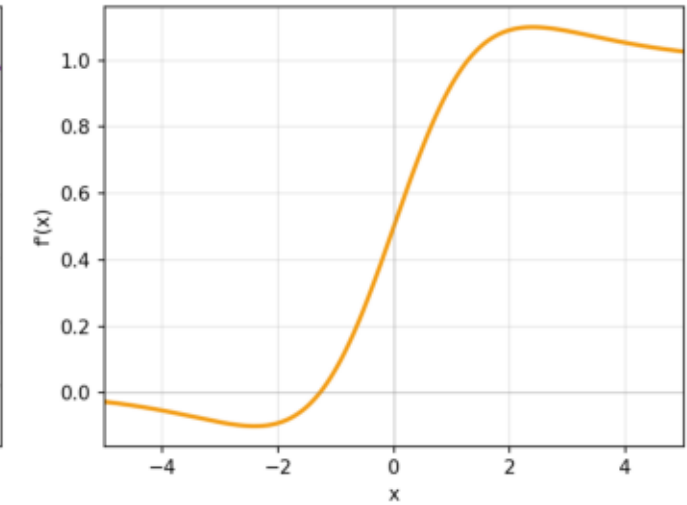
Tanh derivative



GELU derivative

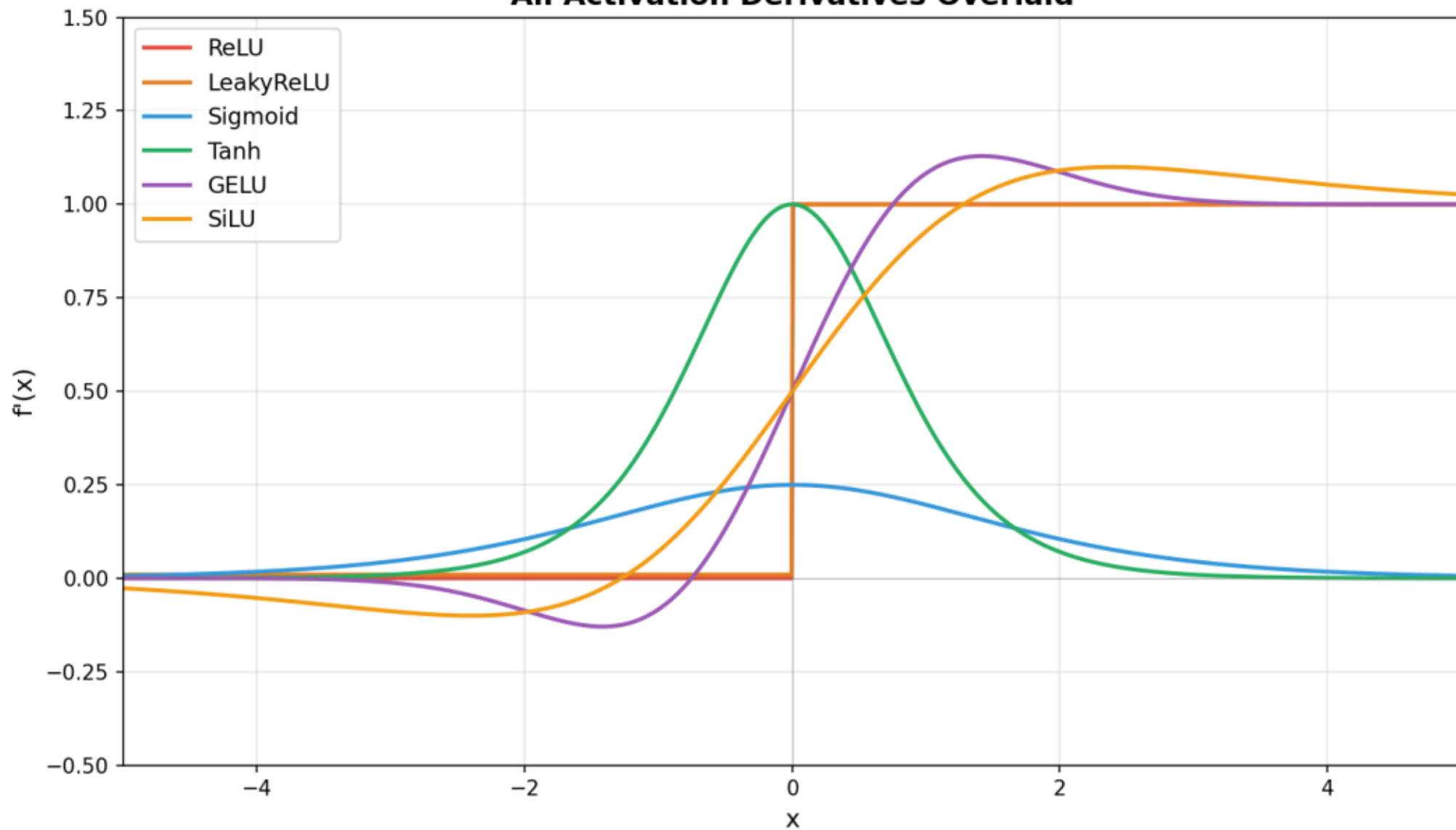


SiLU derivative



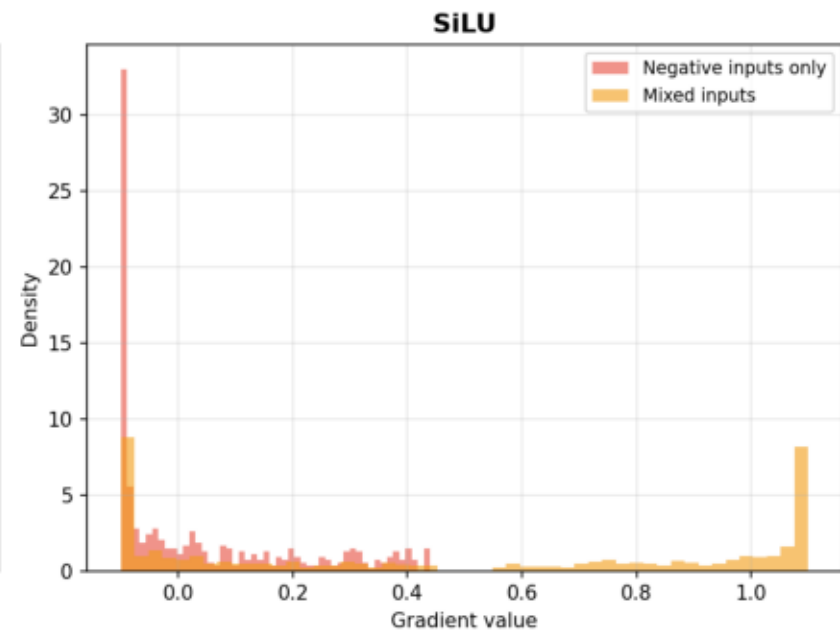
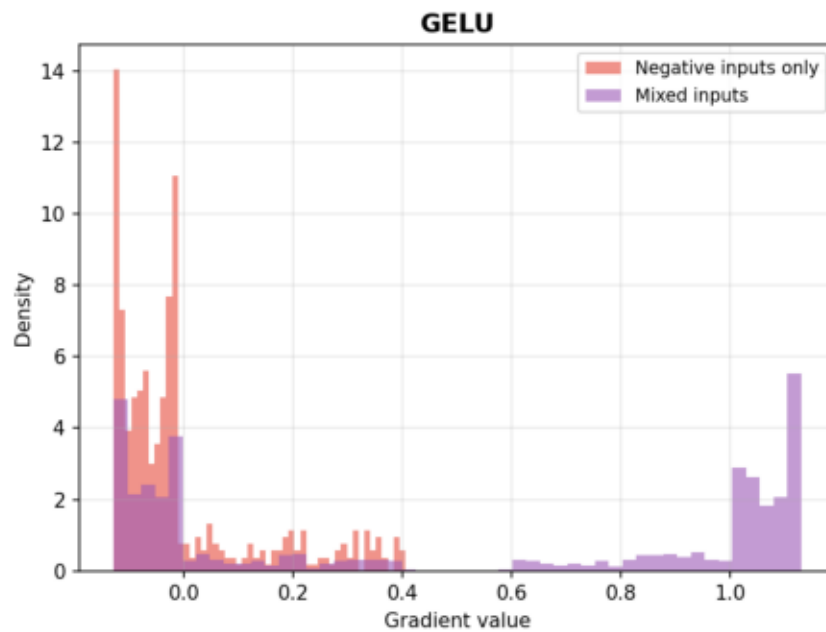
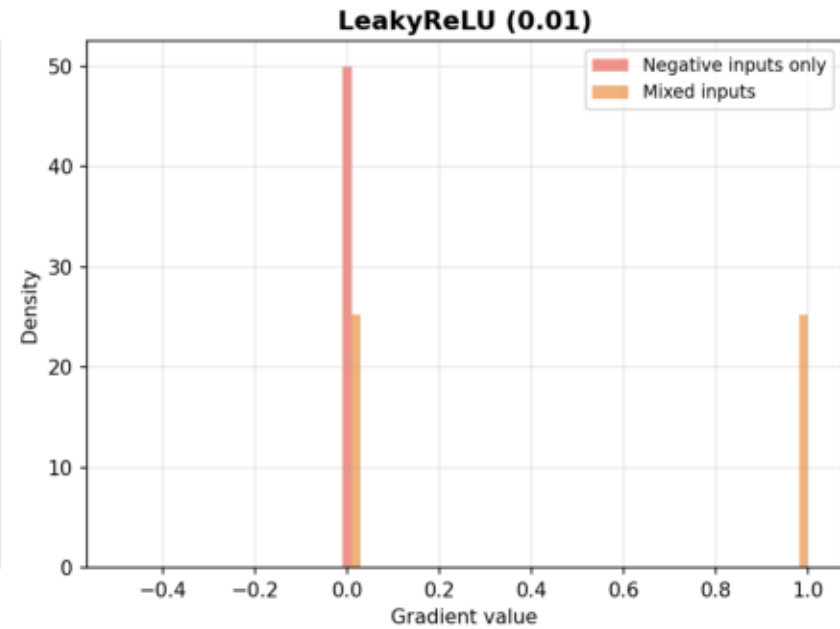
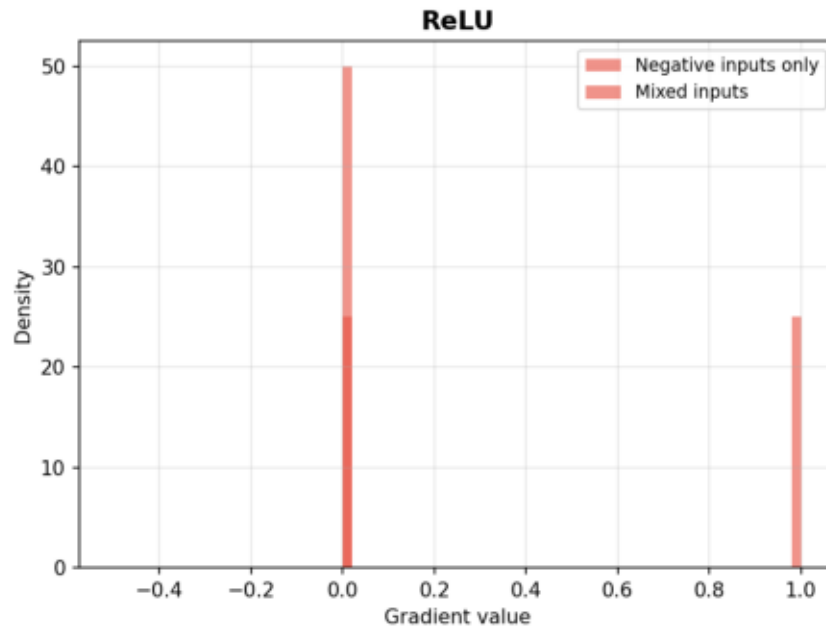
Example 2b: Derivatives (Overlay)

All Activation Derivatives Overlaid



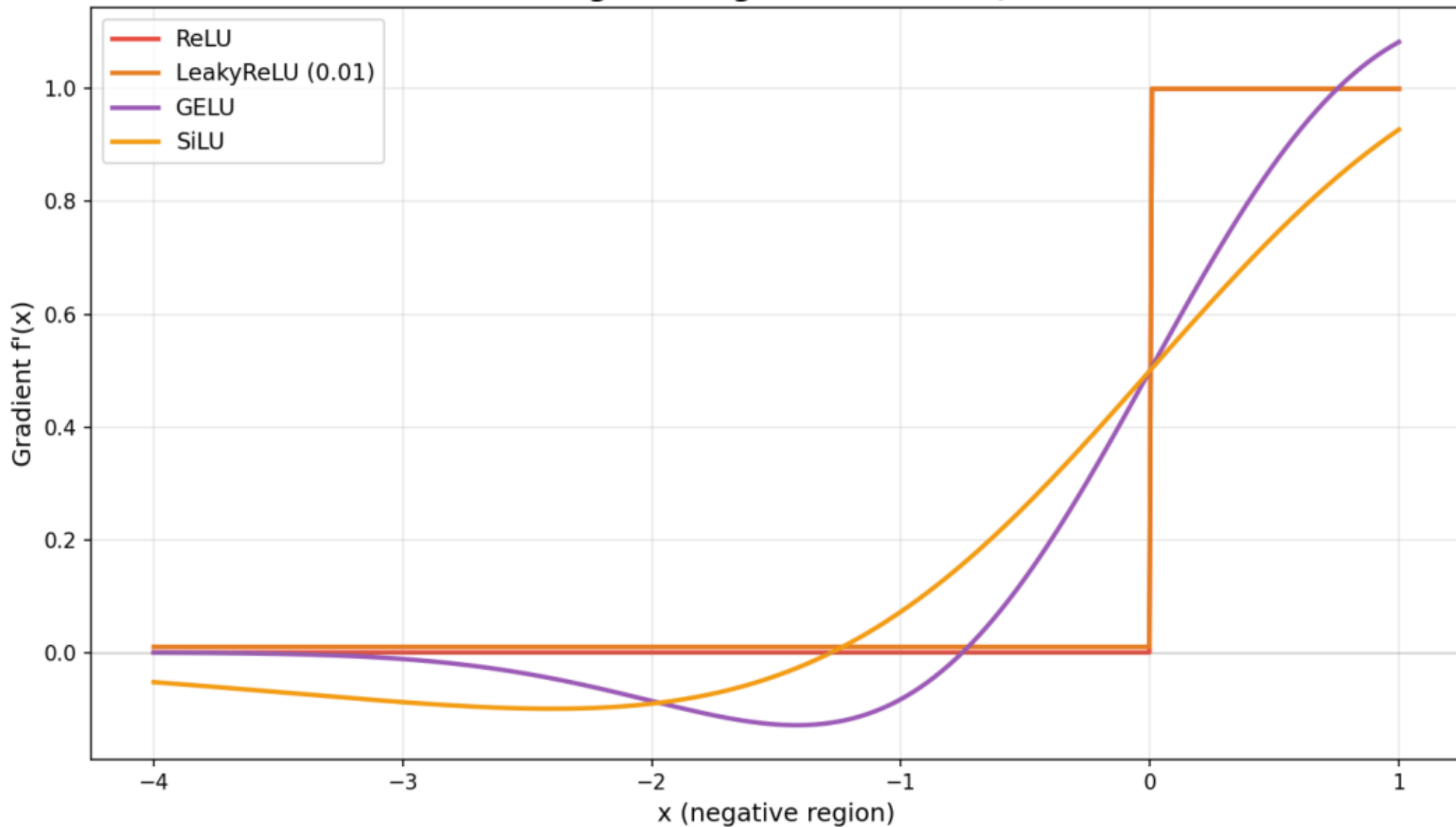
Example 3: Dying ReLU — Gradient Distributions

Dying ReLU Problem — Gradient Distributions



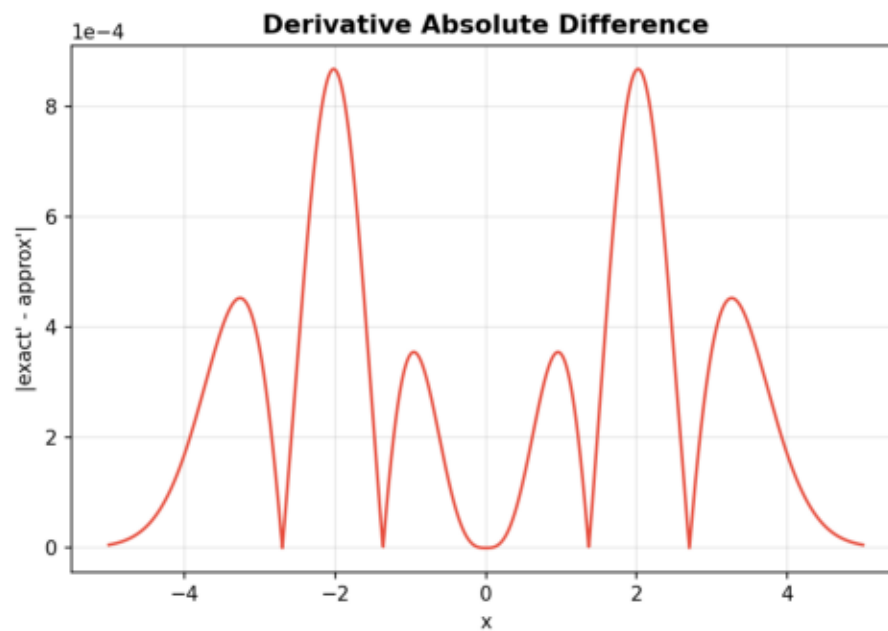
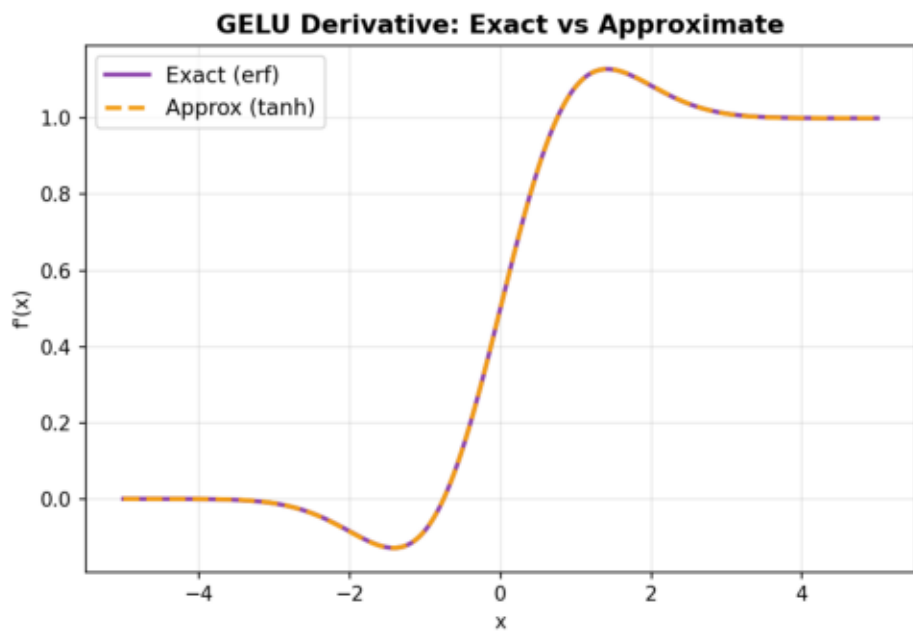
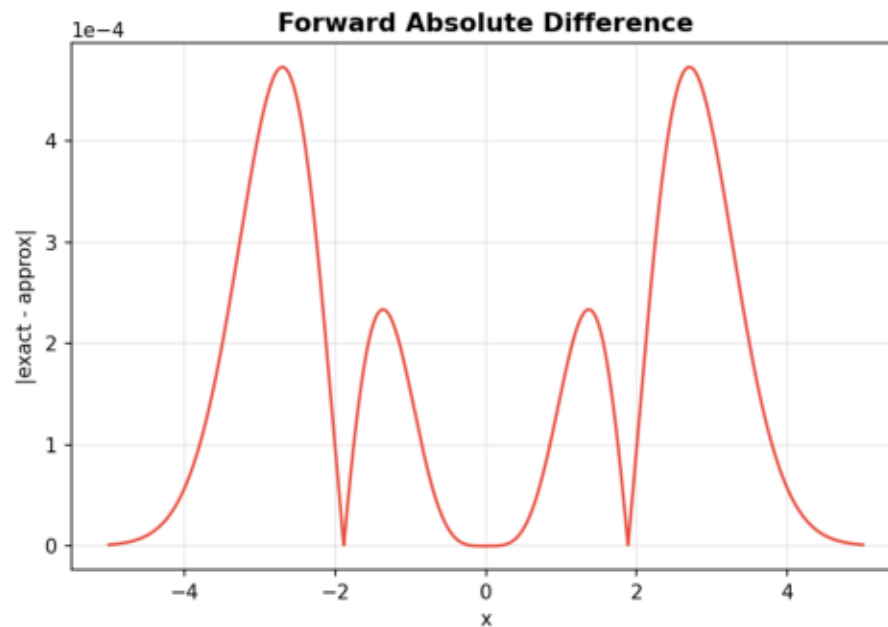
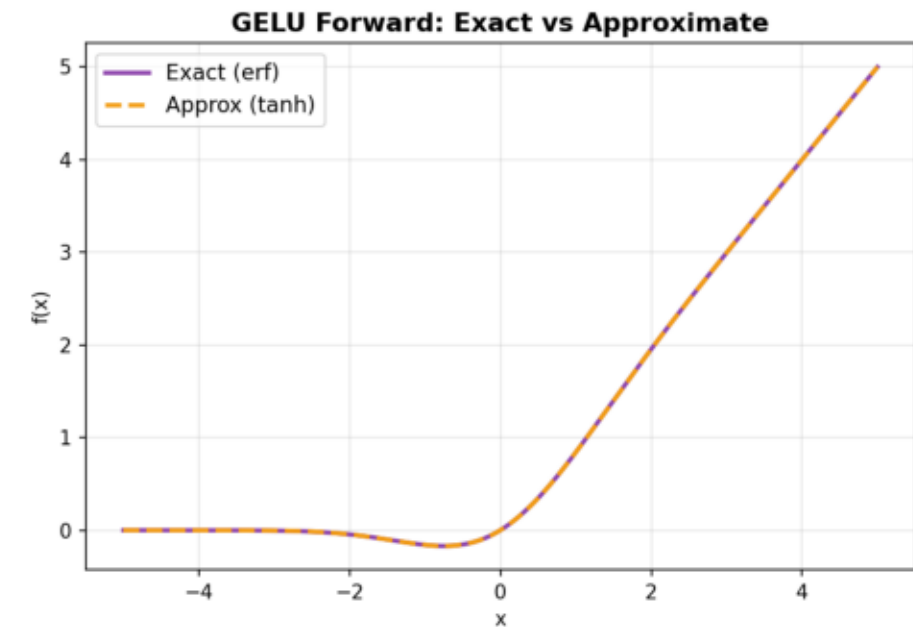
Example 3b: Dying ReLU — Gradient in Negative Region

Gradient in the Negative Region — ReLU Dies, Others Survive



Example 4: GELU Exact vs Approximate

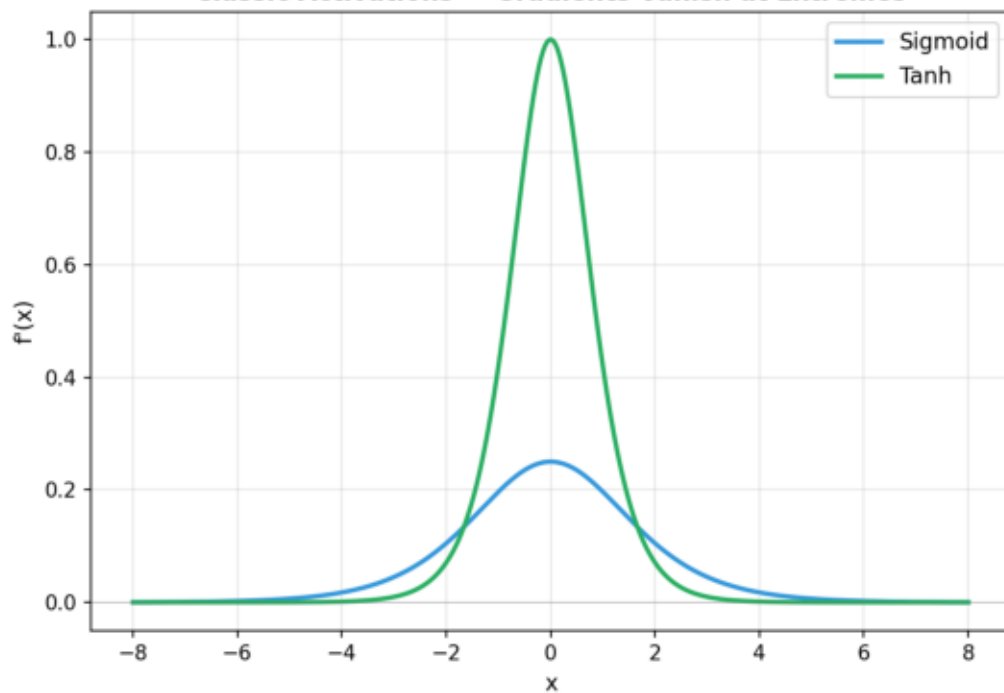
GELU: Exact (erf) vs Tanh Approximation



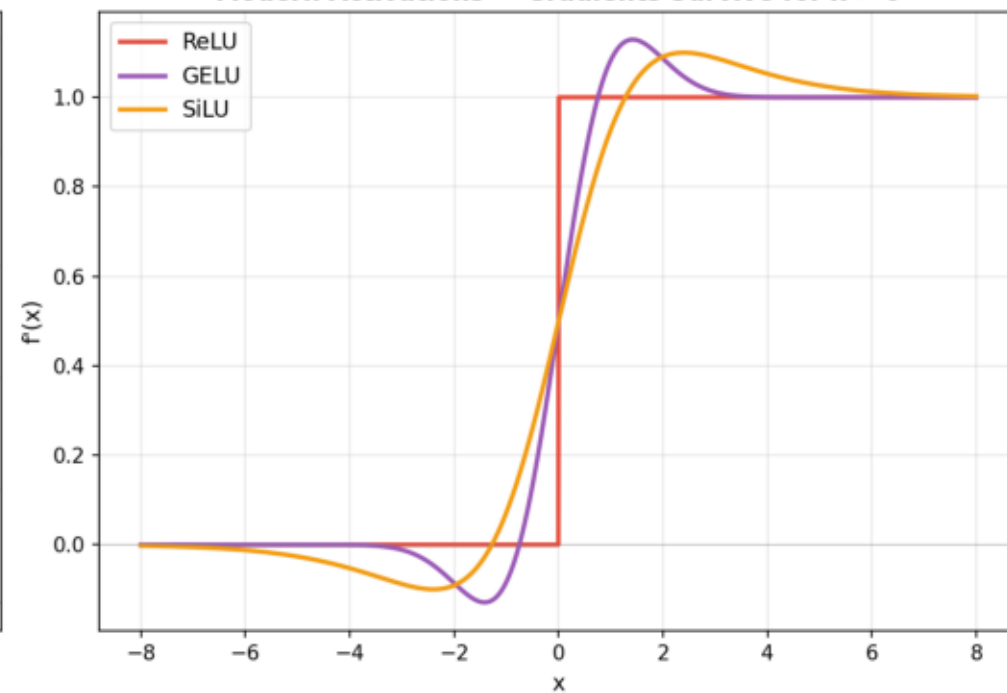
Example 5: Vanishing Gradient Comparison

Vanishing Gradient: Classic vs Modern Activations

Classic Activations — Gradients Vanish at Extremes



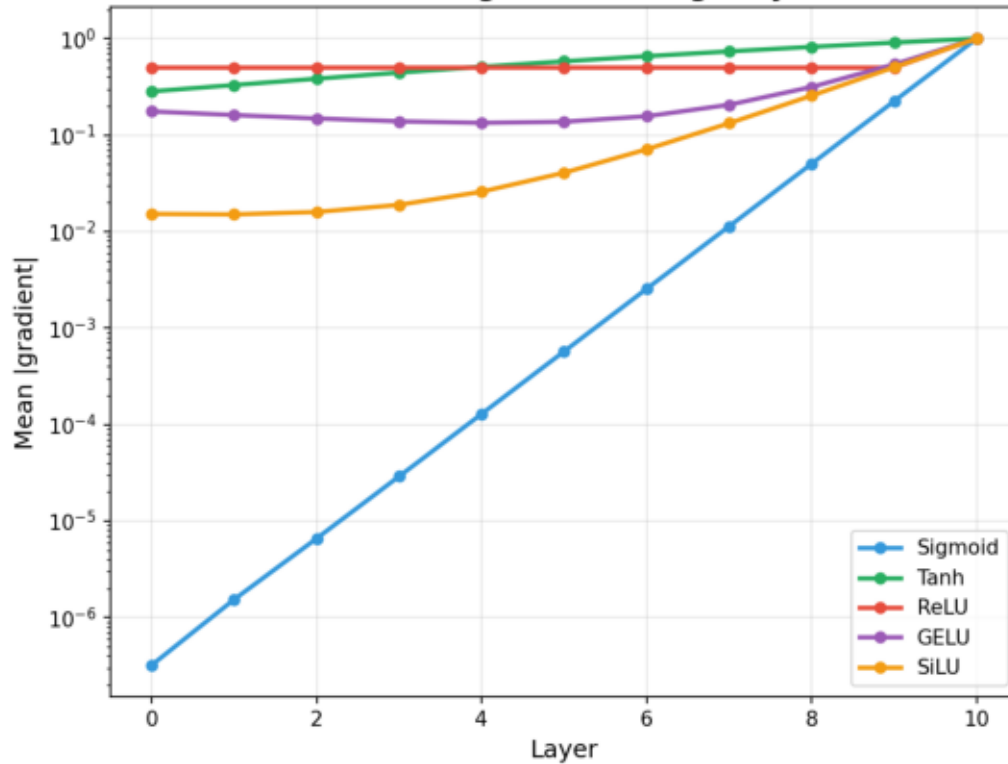
Modern Activations — Gradients Survive for $x > 0$



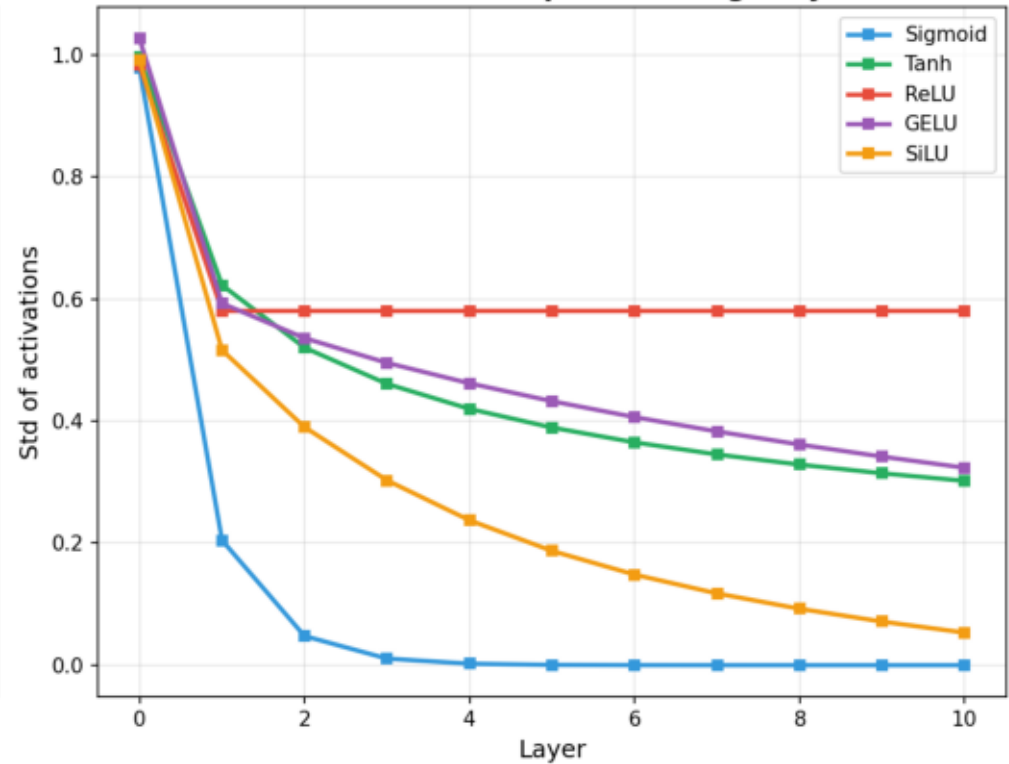
Example 6: Gradient Flow Through 10 Layers

Gradient Flow: Chaining 10 Activation Layers (no weight matrices)

Gradient Magnitude Through Layers

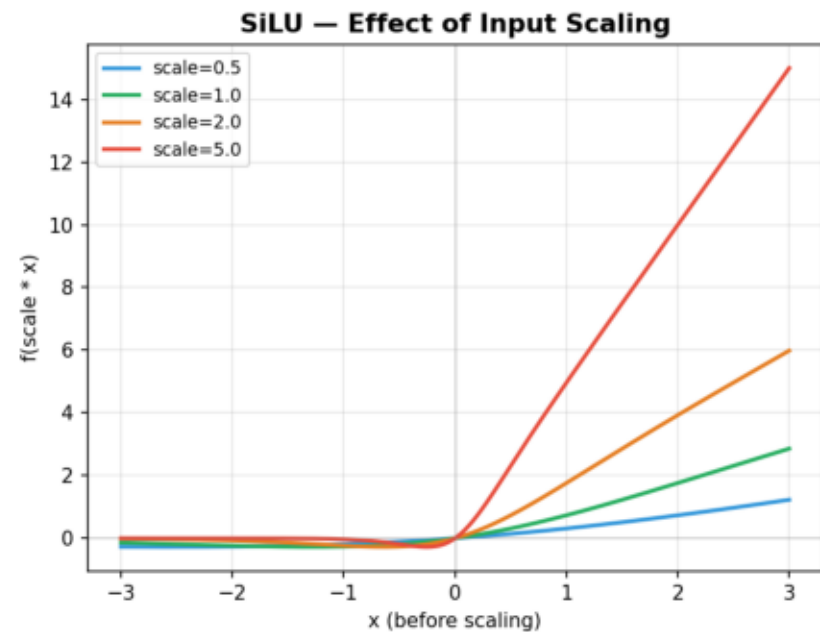
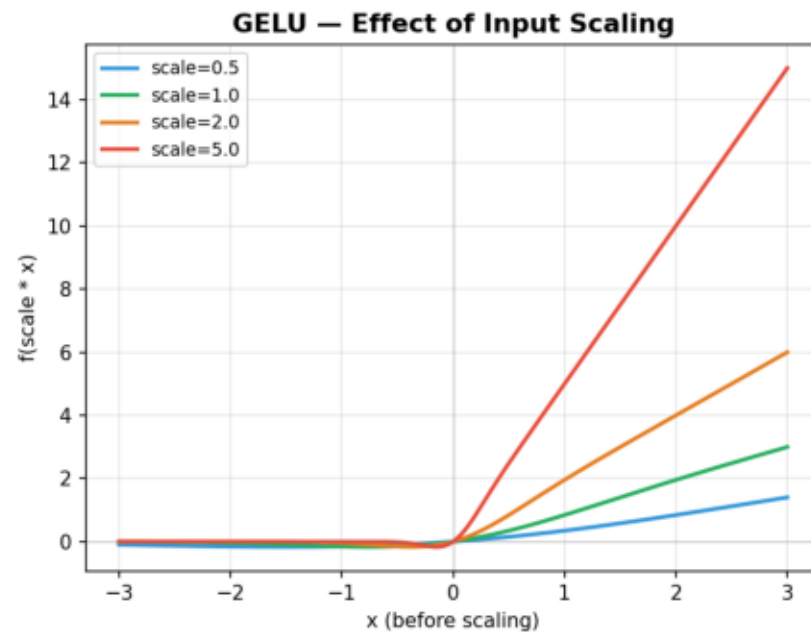
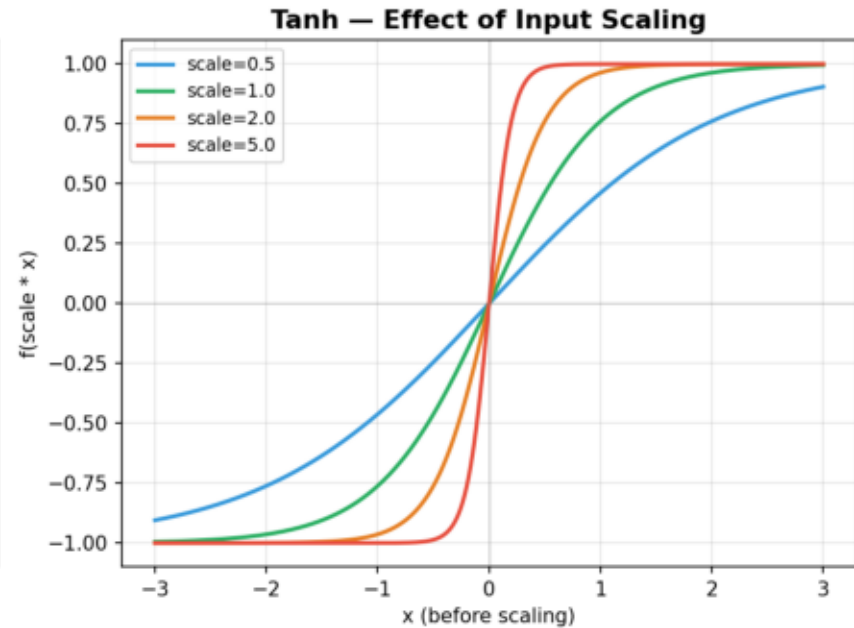
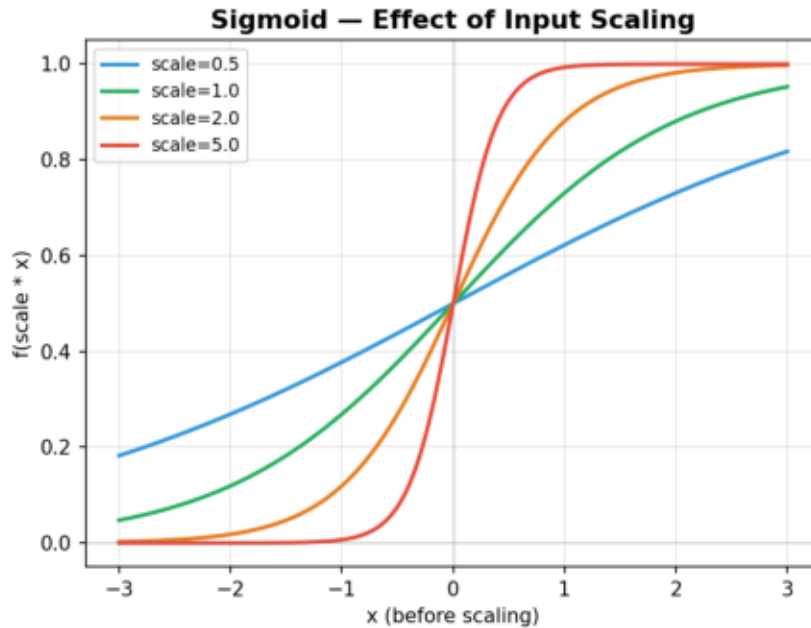


Forward Activation Spread Through Layers



Example 7: Temperature/Scaling Effects

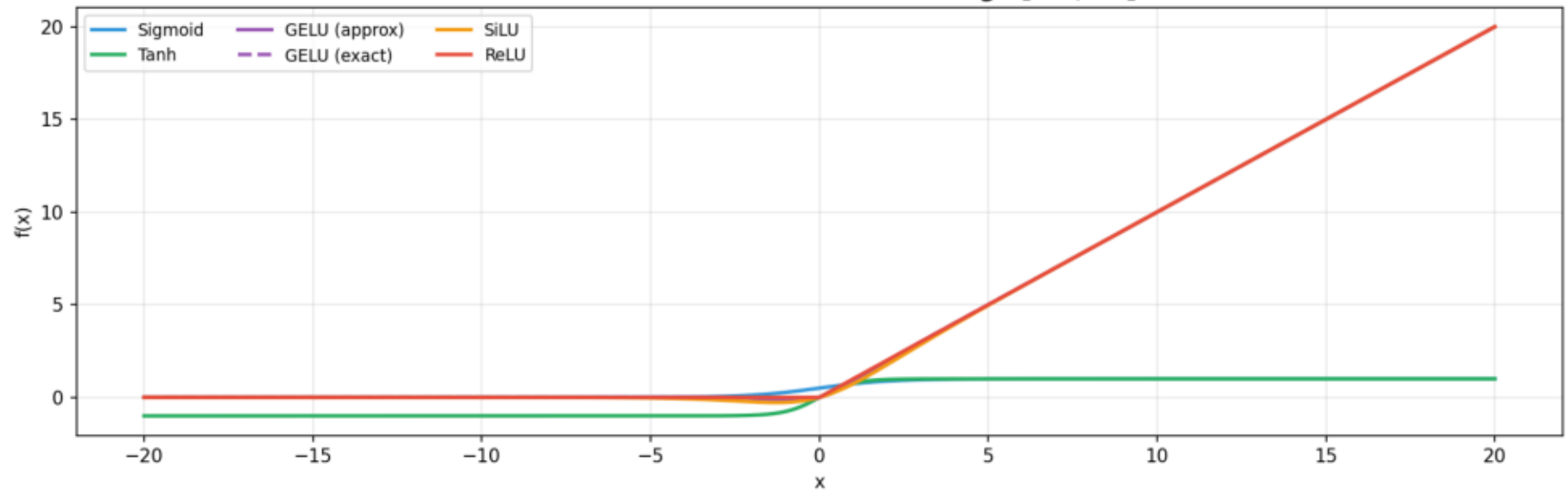
Temperature/Scaling Effects on Activations



Example 8: Numerical Stability

Numerical Stability — Extended Range Behavior

Forward Values Over Extended Range [-20, 20]



Gradient Values Over Extended Range [-20, 20]

