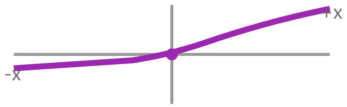
Swish & GELU: Modern State-of-the-Art Activations

Swish

Self-Gating Activation

$f(x) = x \times \sigma(\beta x)$

$\beta = 1$ (Swish-1)



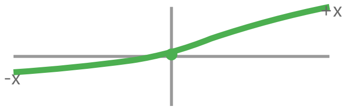
- ✓ Self-gating mechanism
- ✓ Smooth, non-monotonic
- ✓ Neural architecture search

GELU

Gaussian Error Linear Unit

$f(x) = x \times \Phi(x)$

Φ = Gaussian CDF



- ✓ Probabilistic interpretation
- ✓ Smooth, non-monotonic
- ✓ State-of-the-art in NLP

ADOPTION

BERT, GPT (Default)

Key Properties

- ▸ Smooth everywhere
- ▸ Outperform ReLU in Transformers
- ▸ Higher computational cost

Best Use Cases

- 🎯 Transformer models
- 🎯 NLP tasks (BERT, GPT)
- 🎯 State-of-the-art performance