Lecture 21: GELU Activation Function

GELU (Gaussian Error Linear Unit) is an activation function used in Transformers (BERT, GPT, etc.).

Formula:

$$\text{GELU}(x) = x \cdot \Phi(x)$$

where $\Phi(x)$ is the CDF of a standard normal distribution.

Practical approximation:

$$\text{GELU}(x) \approx 0.5x \left(1 + \tanh\left(\sqrt{\frac{2}{\pi}}(x + 0.044715x^3)\right)\right)$$

ReLU says:

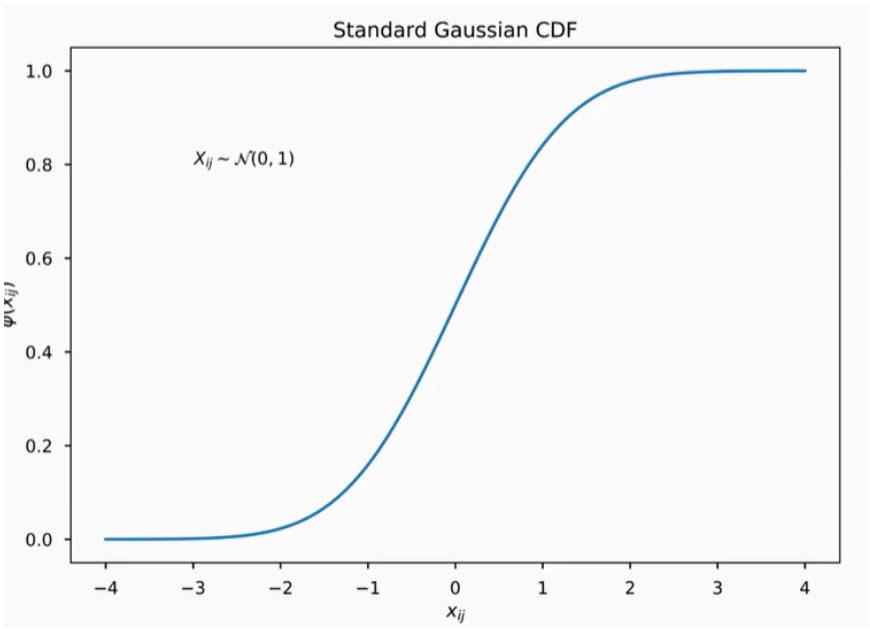
> If x > 0 → keep it, else → zero.

GELU says:

> Keep x, but **scale it smoothly depending on how large it is**.

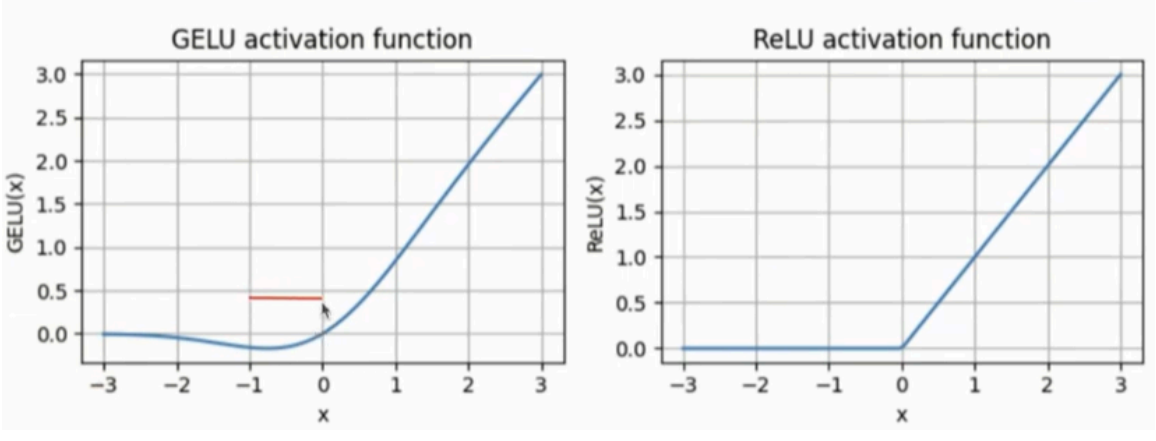
Small negative values aren’t brutally killed.

Small positives aren’t fully trusted either.

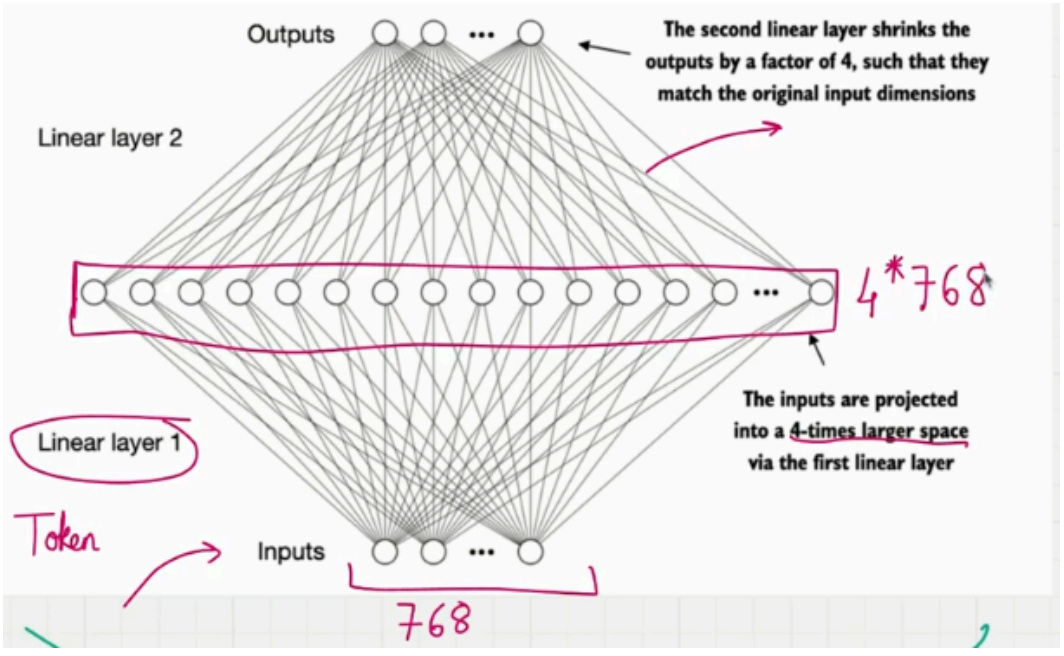


Why LLMs use it

- Smoother than ReLU. In ReLU there is a discontinuity, so it’s not differentiable. But GeLU is differentiable because it’s smooth.
- Better gradient flow
- Empirically improves Transformer performance
- It solves the **Dead Neuron** problem, because it’s not 0 for the negative values of x.



Feed Forward Network



Multi-Head Attention (MHA)	**Feed-Forward Network (FFN)**
Mixes information between tokens	Transforms each token independently
Tokens attend to other tokens	No interaction between tokens
Uses Q, K, V projections and attention weights	Uses Linear → GELU → Linear
Captures relationships and dependencies	Applies nonlinear feature transformation
Computational cost depends on sequence length squared ($O(n^2)$)	Computational cost scales linearly with sequence length ($O(n)$)
Answers: “Which tokens matter?”	Answers: “How should this token be transformed?”