The Effects of Attention Mechanism

Before Attention

All inputs treated equally

	x_1	x_2	x_3	x_4
y_1	0.25	0.25	0.25	0.25
y_2	0.25	0.25	0.25	0.25
y_3	0.25	0.25	0.25	0.25

Static context: Same fixed representation used for all outputs. No selective focus on relevant information.

With Attention

Dynamic attention weights

	x_1	x_2	x_3	x_4
y_1	0.7	0.2	0.05	0.05
y_2	0.1	0.6	0.25	0.05
y_3	0.05	0.15	0.3	0.5

Dynamic context: Different attention distribution for each output. Focuses on most relevant inputs at each step.

🎯 Better Alignment

Learns input-output correspondence automatically. Essential for translation tasks.

📏 Long Sequences

Handles long inputs effectively without information bottleneck. No gradient vanishing.

🔍 Interpretability

Attention weights visualize what model focuses on. Debugging and understanding easier.

Translation Quality	Long Sequences	Gradient Flow
↑ 15-20%	↑ 30-40%	✓ Stable
BLEU Score Improvement	Performance on 50+ tokens	Direct path to each input