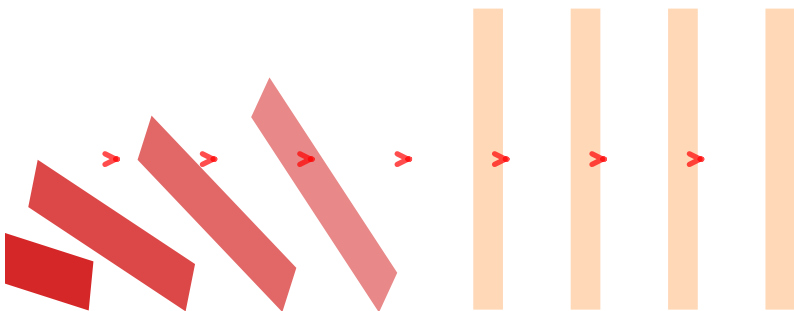


Chart 1: The Domino Effect

RNN: Sequential Processing

Each token waits for the previous one



Transformer: Parallel Processing

All tokens process simultaneously!

