Key Limitations of Standard RNNs

▽ Vanishing Gradient Problem

t_1	→	t_2	→	t_3	→	t_4	→	t_5

- Gradients diminish exponentially during backpropagation
- Early time steps receive negligible weight updates
- Hinders learning from distant past information

↔ Long-term Dependencies

t_1	———————————— weak	t_{10}

Information decays over long sequences

- Difficulty capturing dependencies across long sequences
- Context from early tokens gets lost or diluted
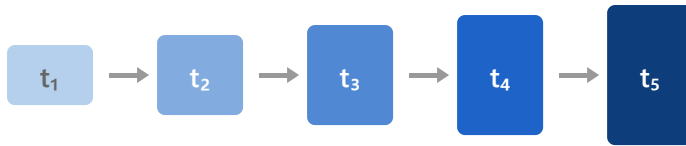- Limits performance on tasks requiring long-range context

↗ Exploding Gradient Problem

⏱ Sequential Processing Bottleneck

t_1	→	t_2	→	t_3	→	t_4

$t_1 \rightarrow t_2 \rightarrow t_3 \rightarrow t_4 \rightarrow t_5$

- Must process sequence step-by-step sequentially

- Cannot leverage parallel computation effectively

- Slow training and inference on long sequences

- Gradients grow exponentially during backpropagation

- Causes numerical instability and training divergence

- Requires gradient clipping to stabilize training