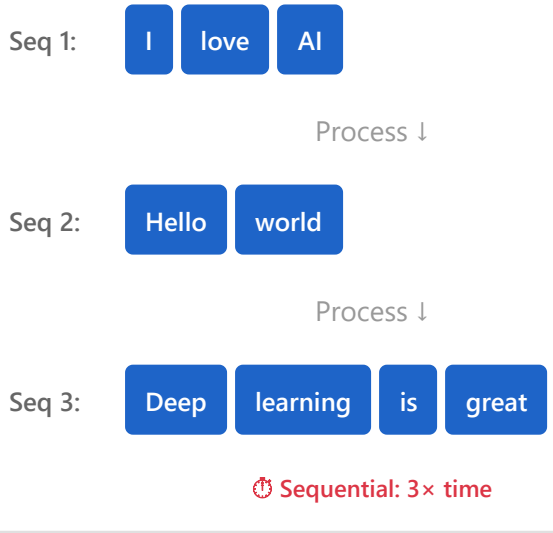
Batching in Sequence Models

⚠ Challenge: Variable-Length Sequences

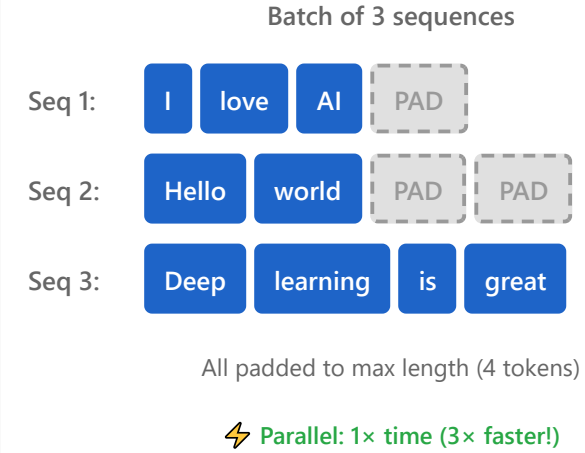
Sequences in a batch often have different lengths. We need to pad them to the same length for efficient parallel processing on GPUs.

Without Batching



Sequential processing: Process one sequence at a time. Slow and inefficient. Cannot leverage GPU parallelism.

With Batching (Padded)



Parallel processing: Process all sequences simultaneously. Fast and efficient. Full GPU utilization. Need masking!

Padding Strategy

Add special PAD tokens to shorter sequences to match the longest sequence in batch.

```
pad_sequence(sequences,
  batch_first=True,
  padding_value=0)
```

Use Masking

Create masks to ignore PAD tokens during attention computation and loss calculation.

```
mask = (input != PAD)
scores.masked_fill(
  ~mask, -inf)
```

Batch Size

Balance between memory and speed. Larger batches = faster but more memory. Typical: 32-128.

```
# Common sizes
batch_size = 32
batch_size = 64
batch_size = 128
```