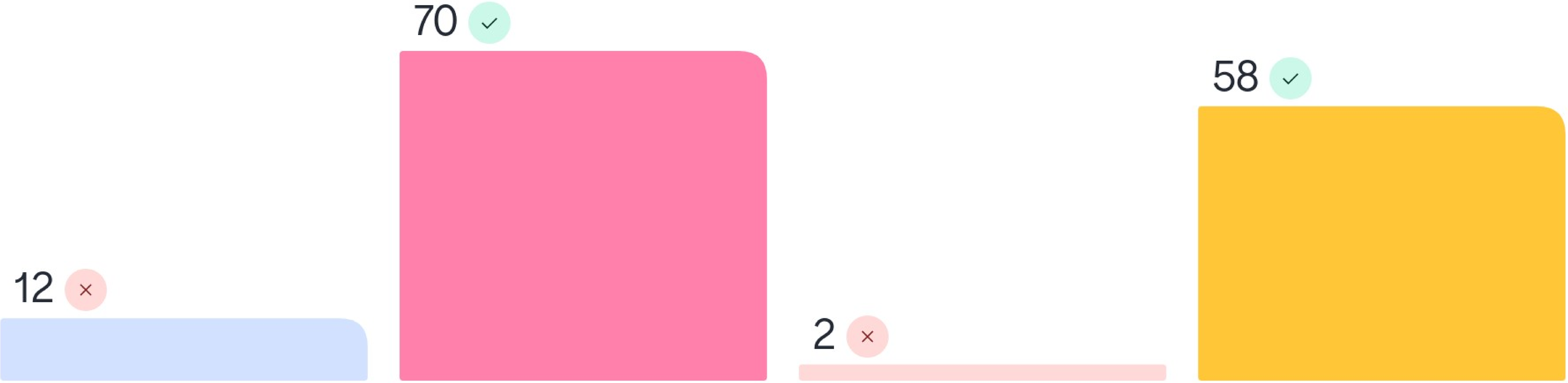
Which of the following are advantages of a Transformer architecture over a recurrent one?



Transformers can process arbitrary length sequences.

During training, Transformers optimize all sequence positions in parallel, unlike RNNs.

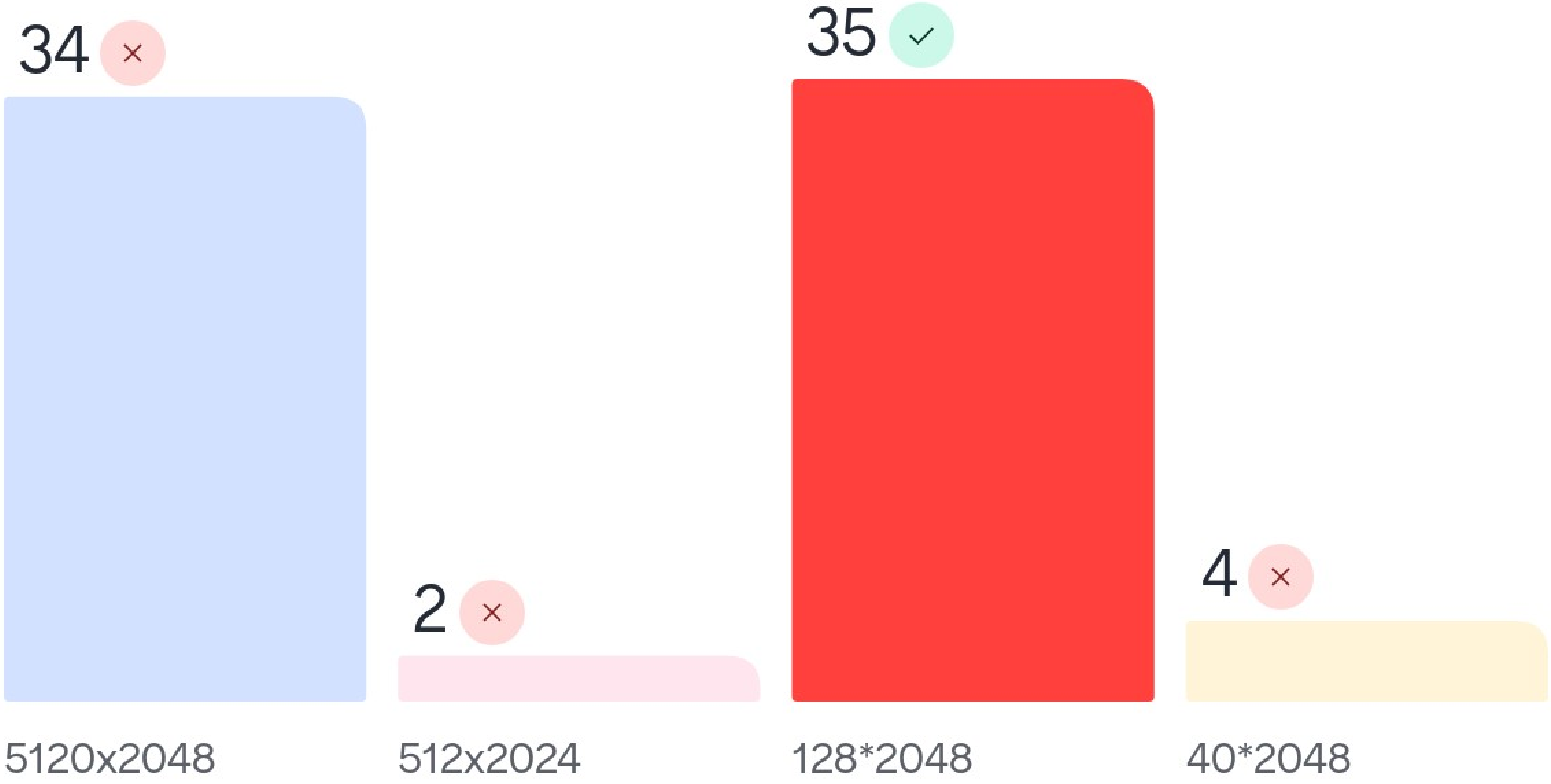
The log likelihood loss we use to optimize Transformers is more numerically stable than the loss used for for RNNs.

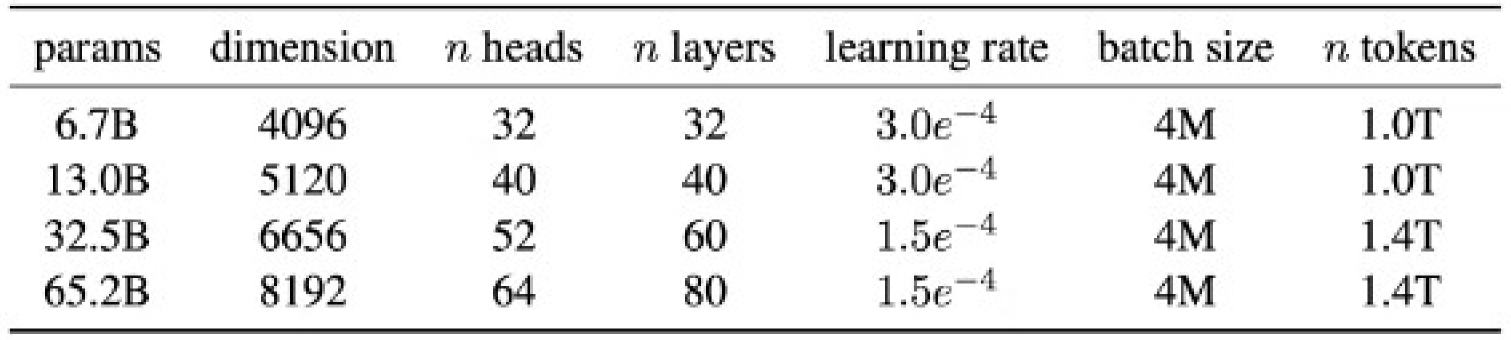
RNNs "forget" information found early in an input sequence; Transformers do not.

7

74

The table on the right gives architectures info for LLaMA. For LLaMA 13B, what was the dimension of the matrices used in the attention computation?





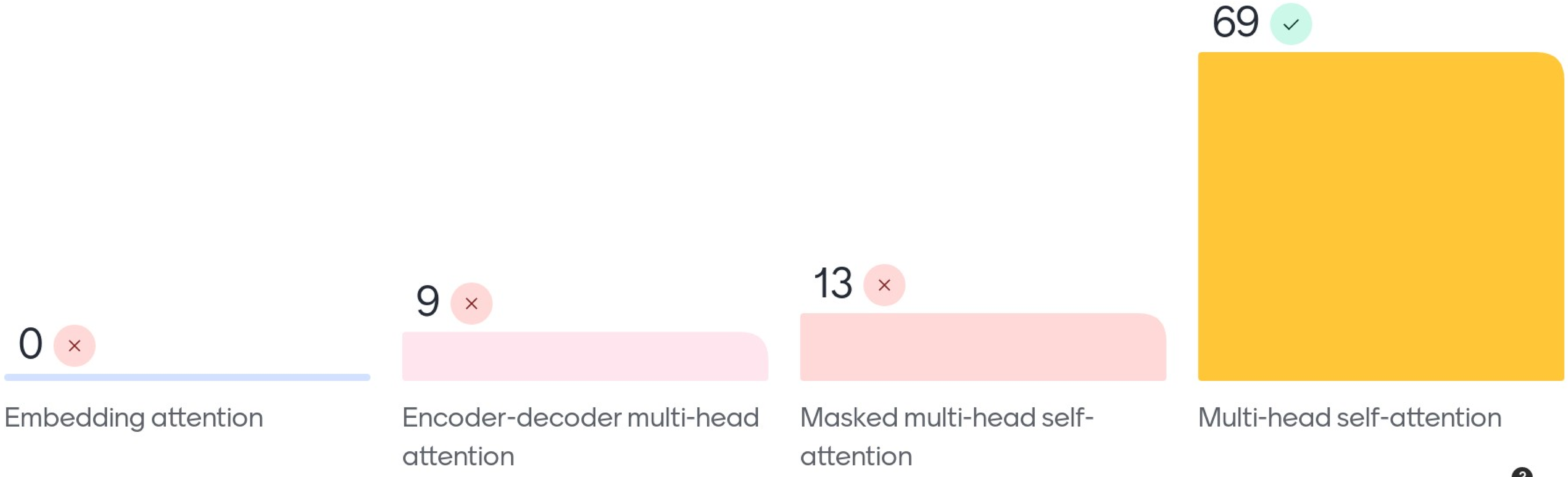
params	dimension	n heads	n layers	learning rate	batch size	n tokens
6.7B	4096	32	32	$3.0e^{-4}$	4M	1.0T
13.0B	5120	40	40	$3.0e^{-4}$	4M	1.0T
32.5B	6656	52	60	$1.5e^{-4}$	4M	1.4T
65.2B	8192	64	80	$1.5e^{-4}$	4M	1.4T

Table 2: **Model sizes, architectures, and optimization hyper-parameters.**

4 👍

64 👤

Which type of attention do you find in the encoder of an encoder-decoder Transformer model?


0 ✗
9 ✗
13 ✗
69 ✓
Embedding attention
Encoder-decoder multi-head attention
Masked multi-head self-attention
Multi-head self-attention

2
75