Attention is all you need

Ashish Vaswani (Google Brain)
Noam Shazeer (Google Brain)
Niki Parmar (Google Research)
Jakob Uszkoreit (Google Research)
Llion Jones (Google Research)
Aidan N. Gomez † (University of Toronto)
Łukasz Kaiser (Google Brain)
Illia Polosukhin

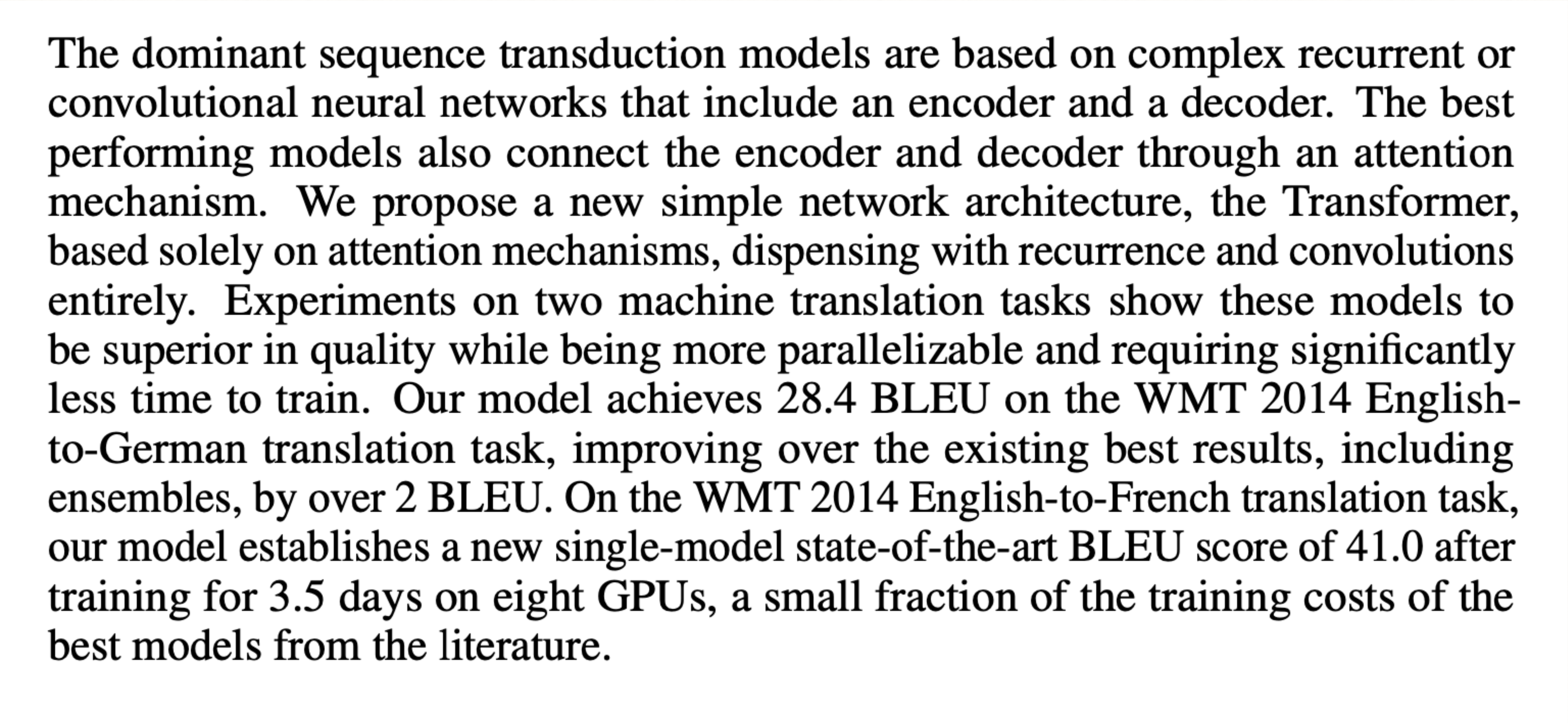
Meta Data & Stats

Published in:	ACM NIPS’17
Year:	2017
Number of Authors:	8
Citations:	59334
Pages (PDF):	10
Figures:	2
References:	36
Formals:	0 definitions

Table of Content

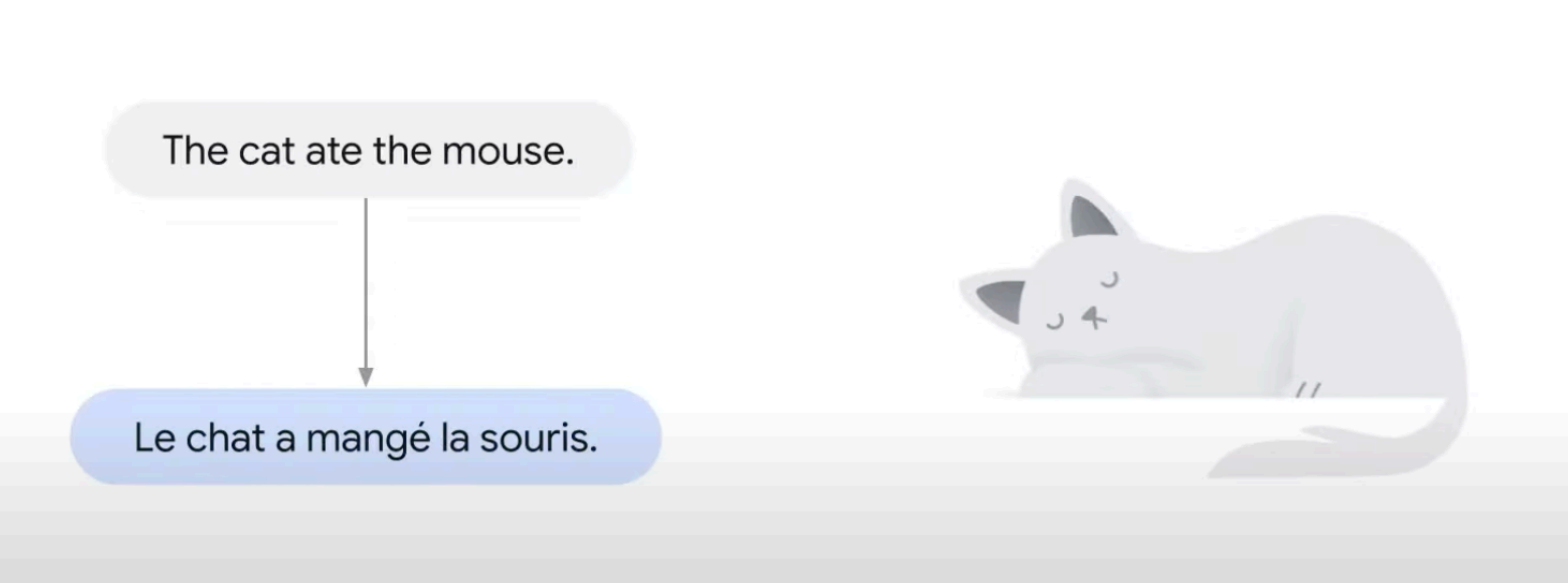
1. Introduction
2. Background
3. Model Architecture
4. Why Self-Attention
5. Training
6. Results
7. Conclusion
8. References

Abstract



The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

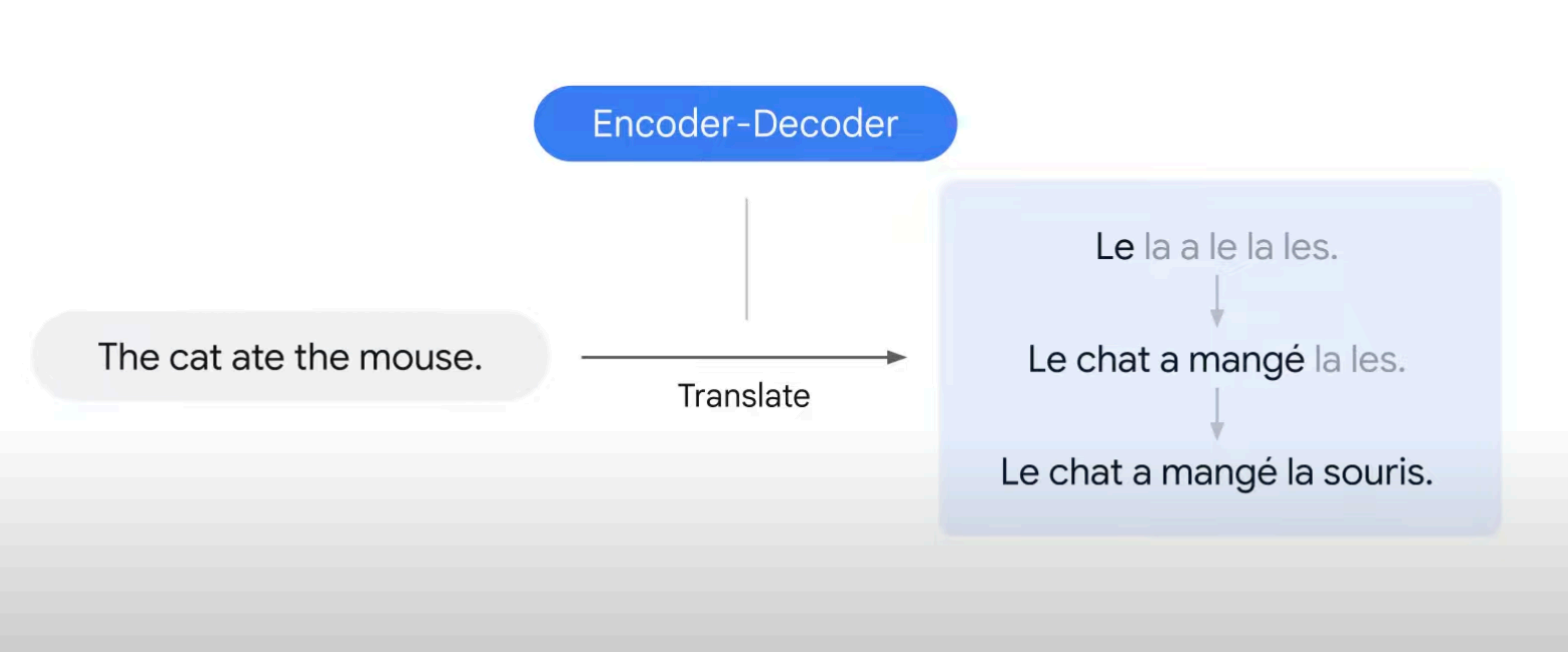
Background



The cat ate the mouse.

Le chat a mangé la souris.

Background



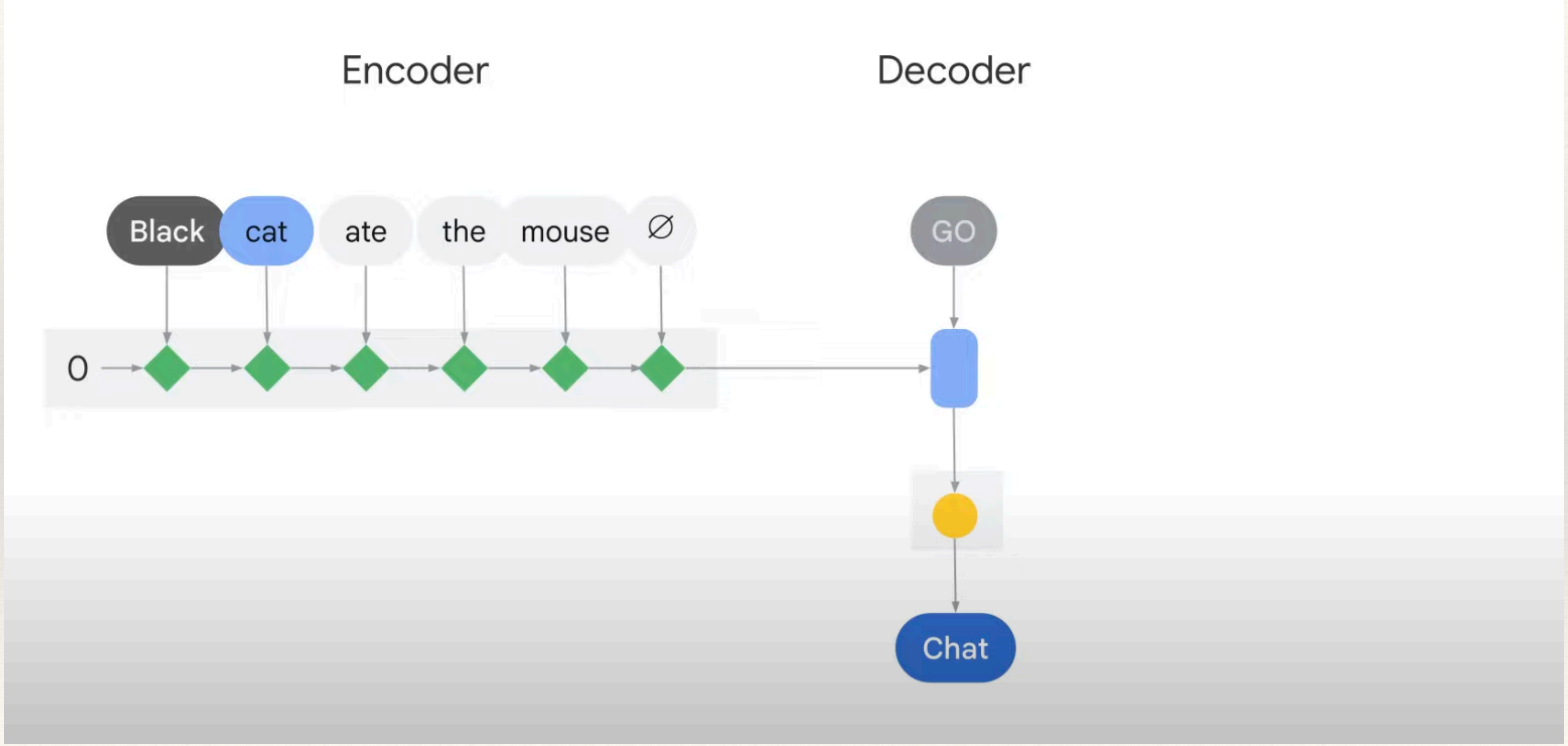
The cat ate the mouse.

Le la a le la les.

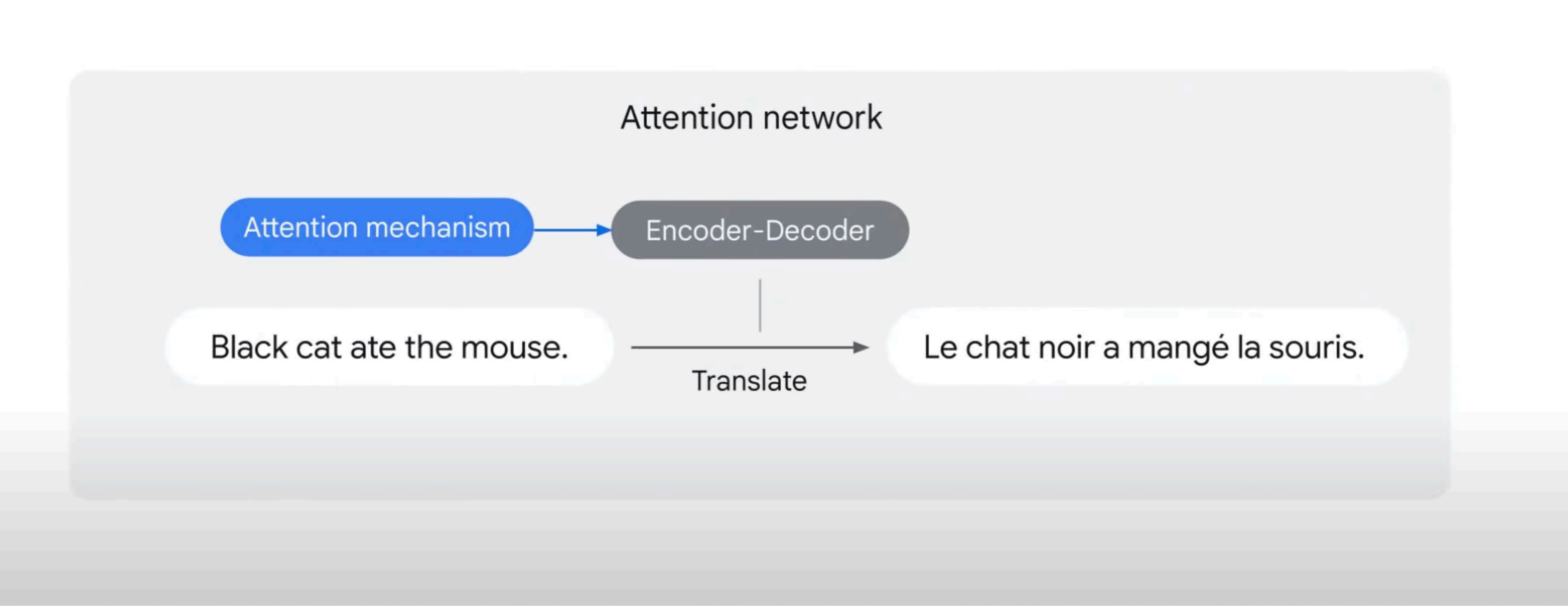
Le chat a mangé la les.

Le chat a mangé la souris.

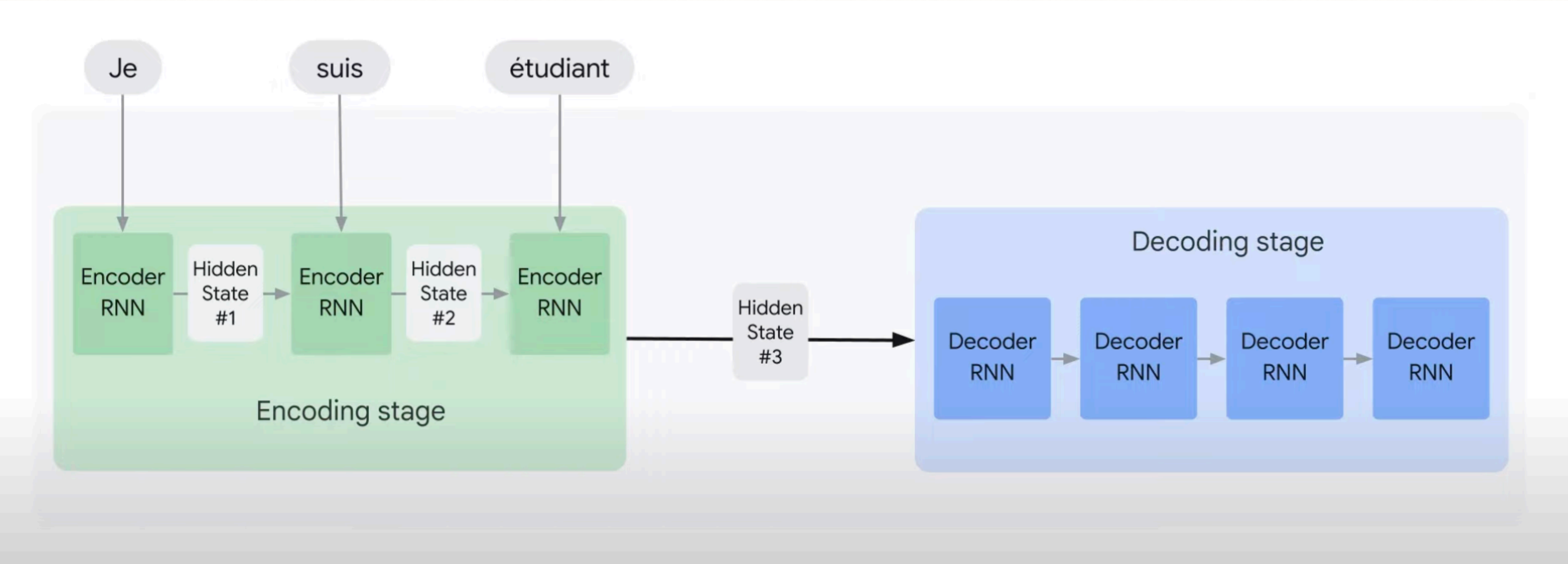
Background



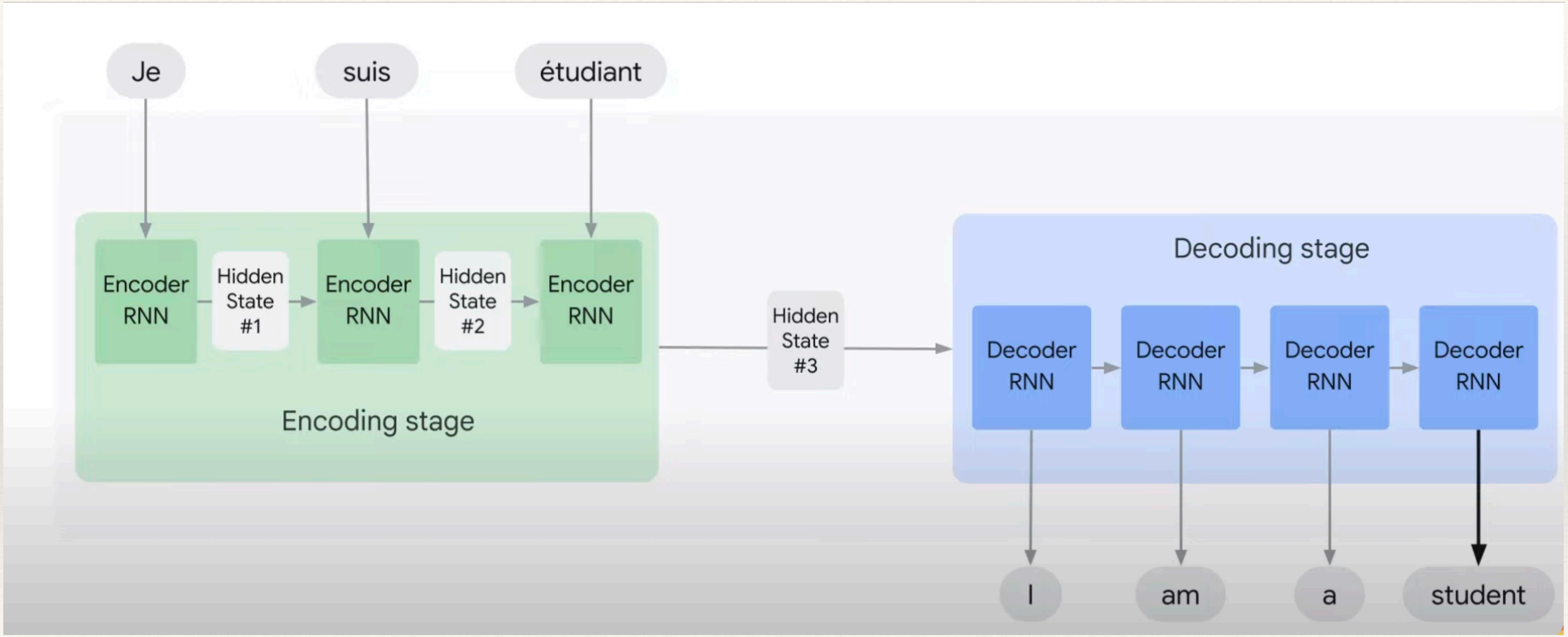
Translation model



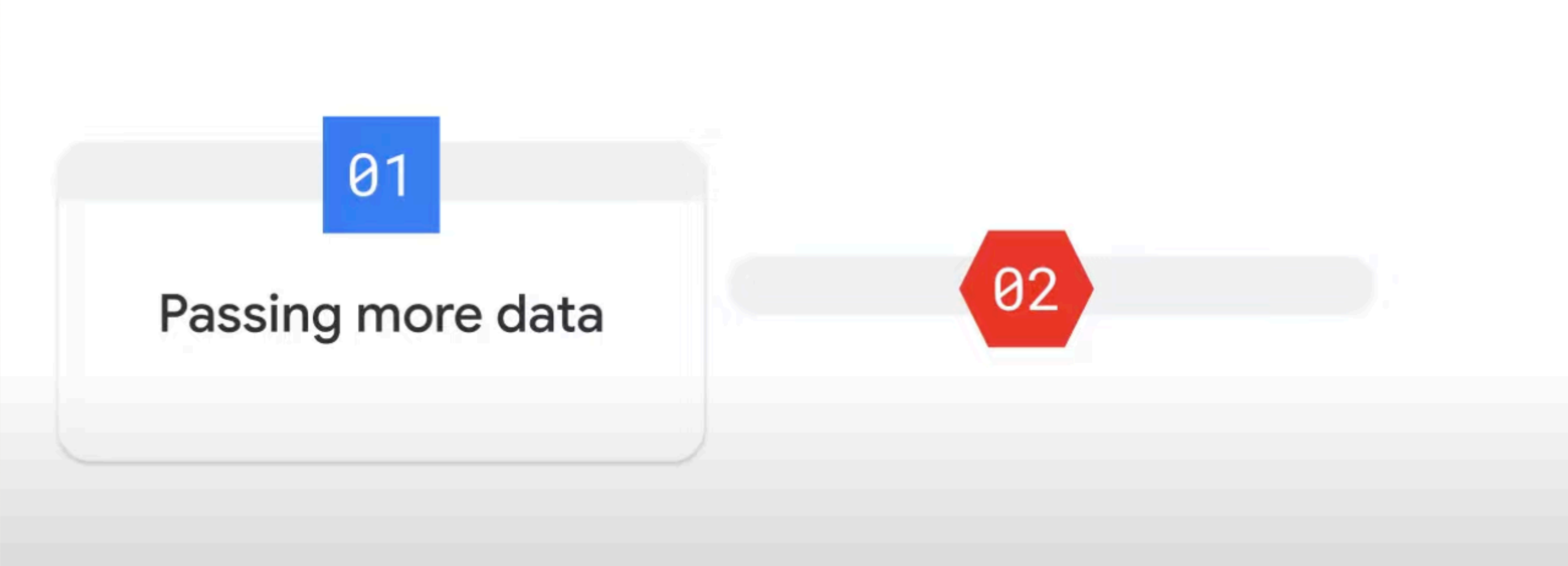
Traditional RNN



Traditional RNN



Attention model differs from a traditional one

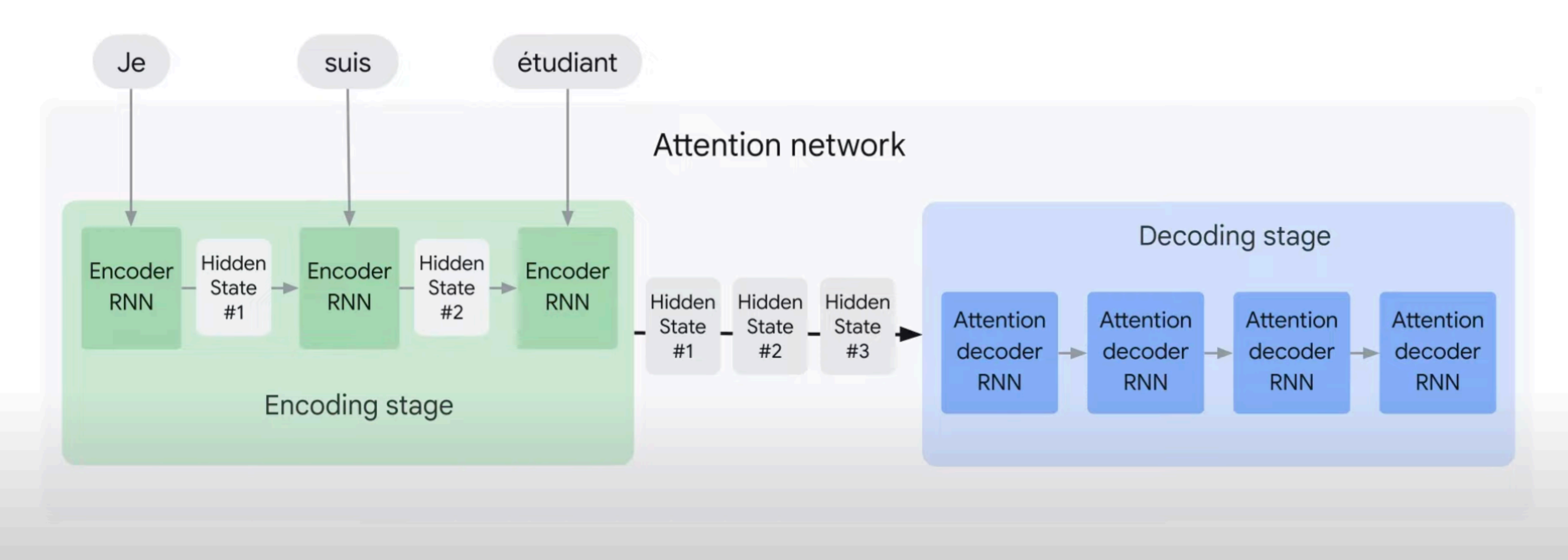


01

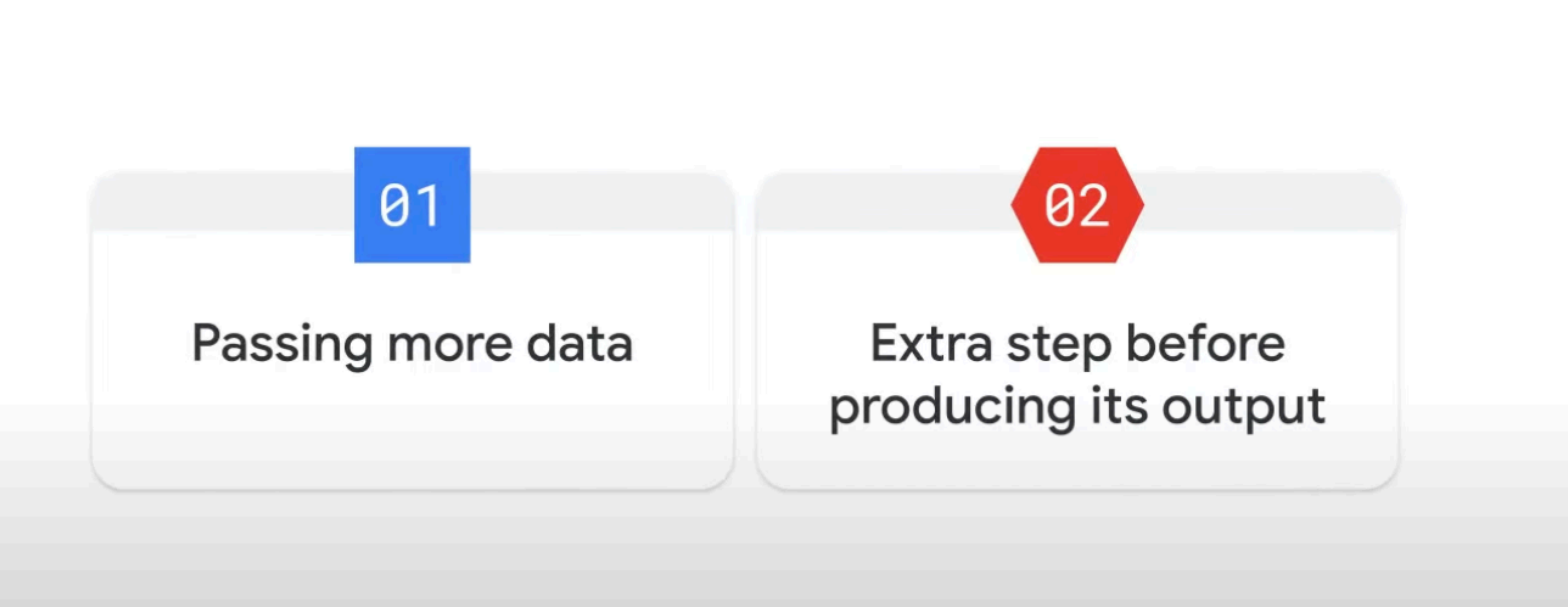
Passing more data

02

Attention model differs from a traditional one



Attention model differs from a traditional one



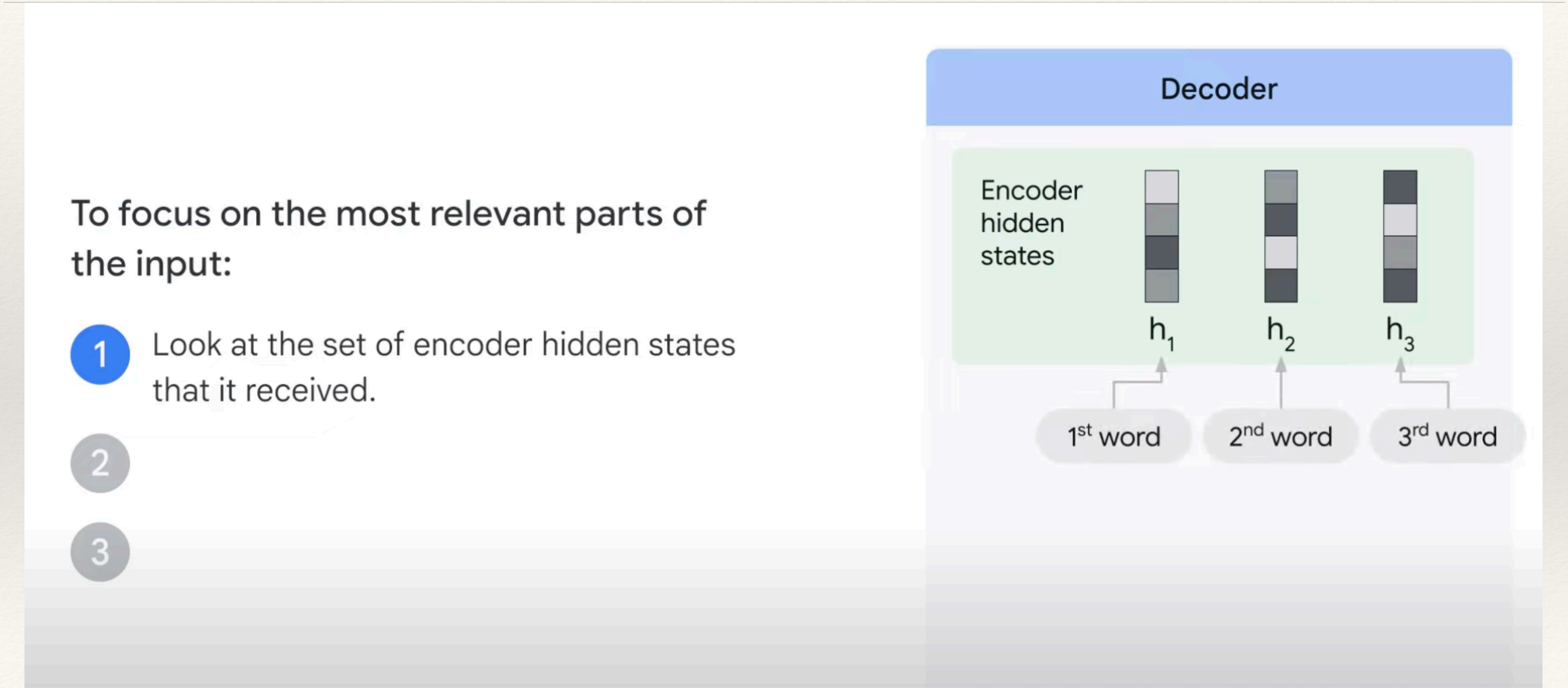
01

Passing more data

02

Extra step before producing its output

Attention model differs from a traditional one



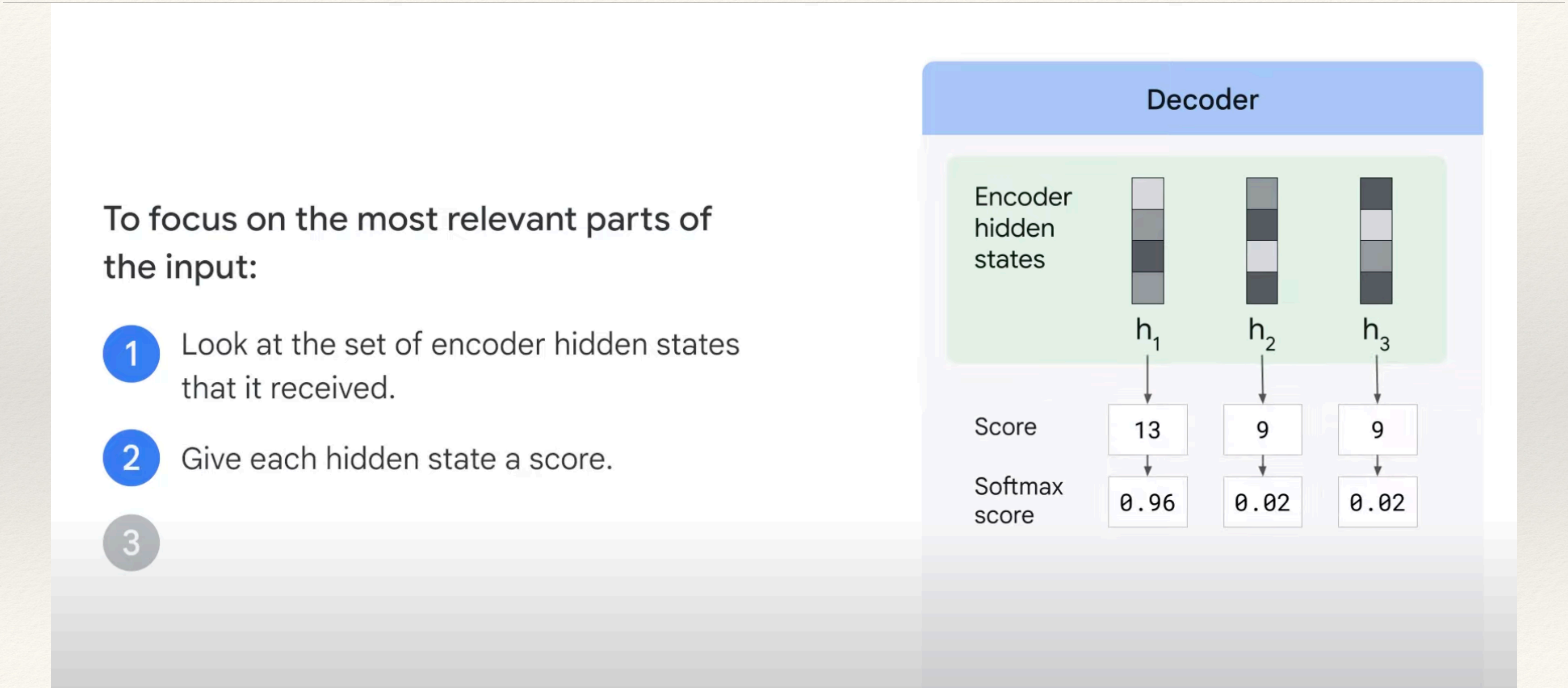
To focus on the most relevant parts of the input:

1 Look at the set of encoder hidden states that it received.

2

3

Attention model differs from a traditional one

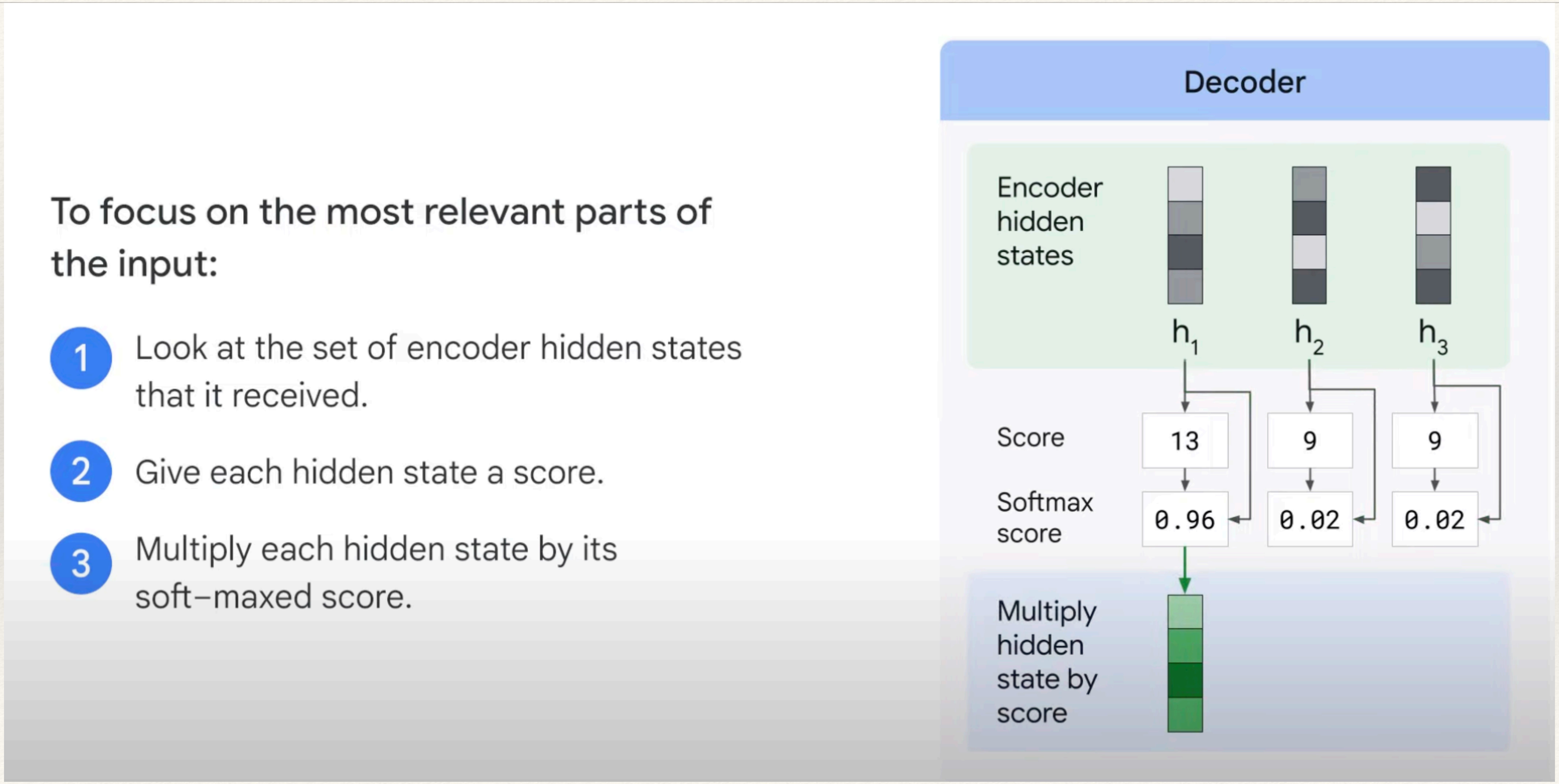


To focus on the most relevant parts of the input:

1. Look at the set of encoder hidden states that it received.
2. Give each hidden state a score.
3.

	h_1	h_2	h_3
Score	13	9	9
Softmax score	0.96	0.02	0.02

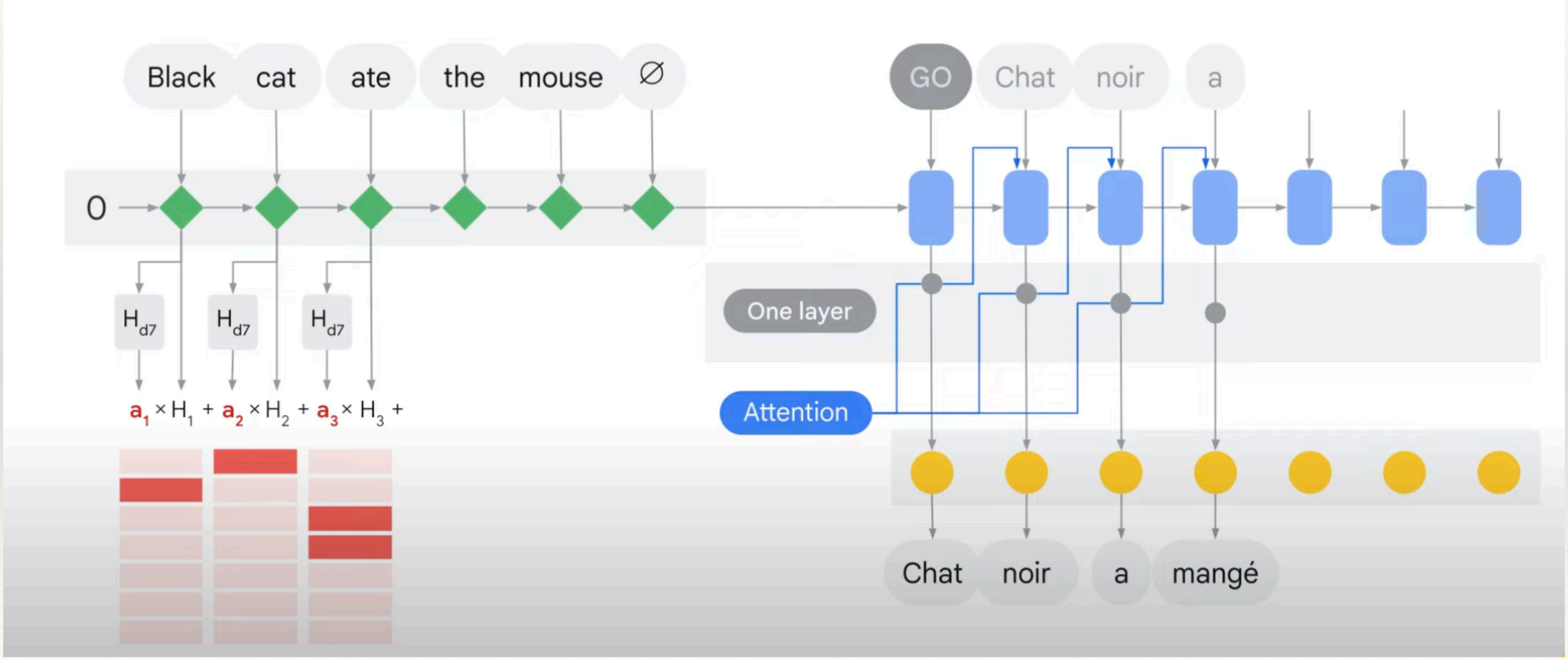
Attention model differs from a traditional one



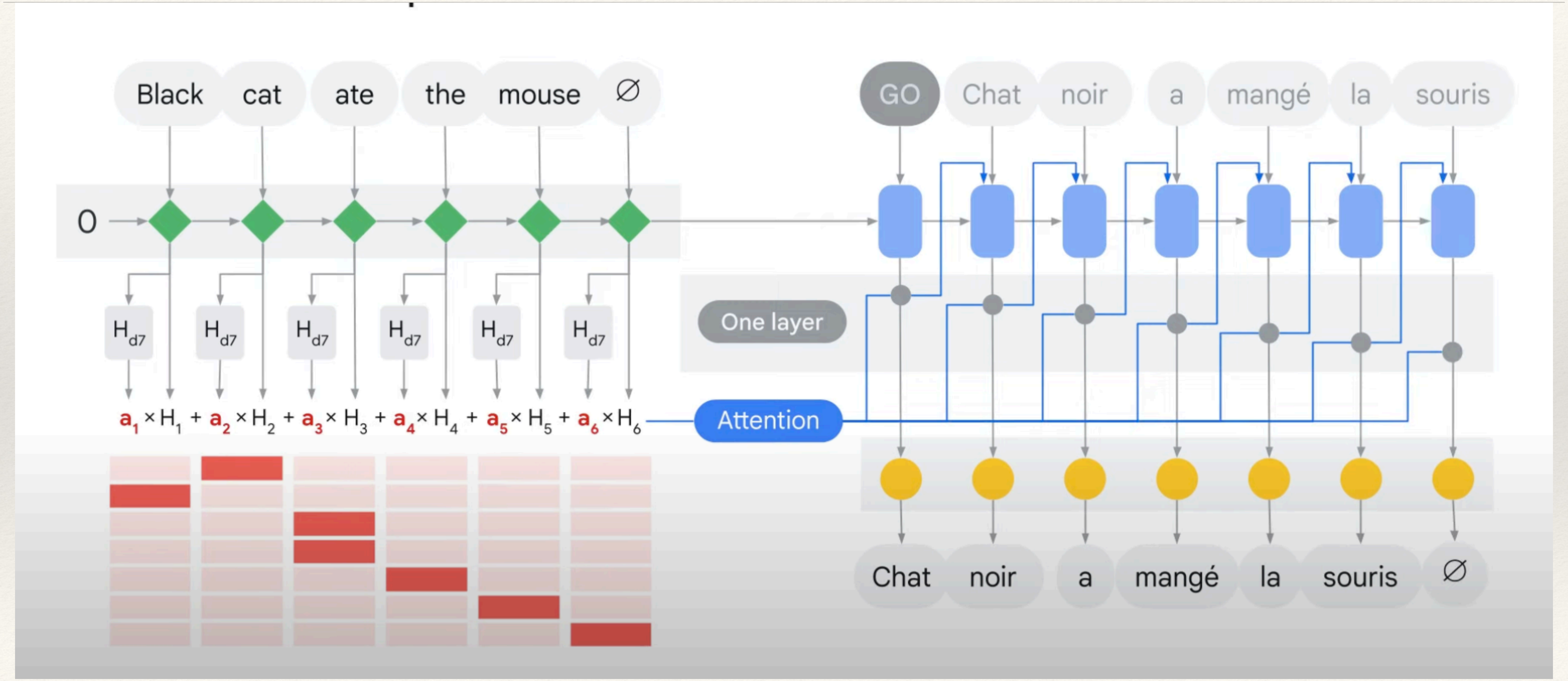
To focus on the most relevant parts of the input:

1. Look at the set of encoder hidden states that it received.
2. Give each hidden state a score.
3. Multiply each hidden state by its soft–maxed score.

Improving translation



Improving translation



Paper

8. Conclusion