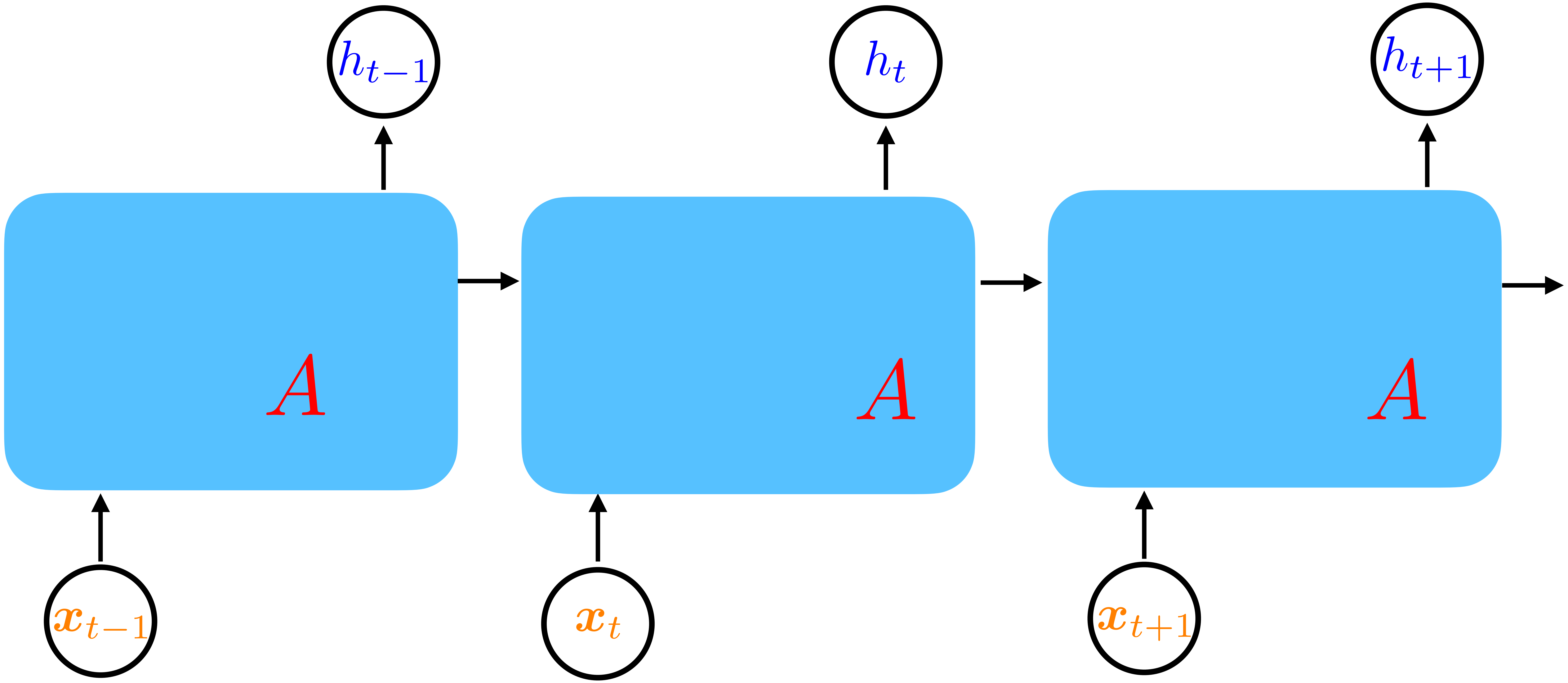
Sequence to Sequence Model and Attention

COMP7607 — Lecture 3

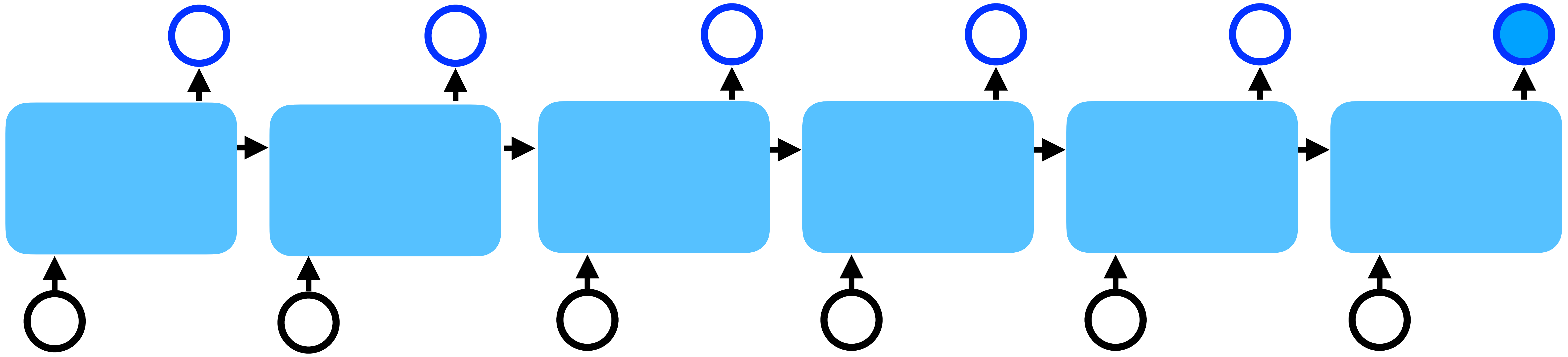
Lingpeng Kong

Department of Computer Science, The University of Hong Kong

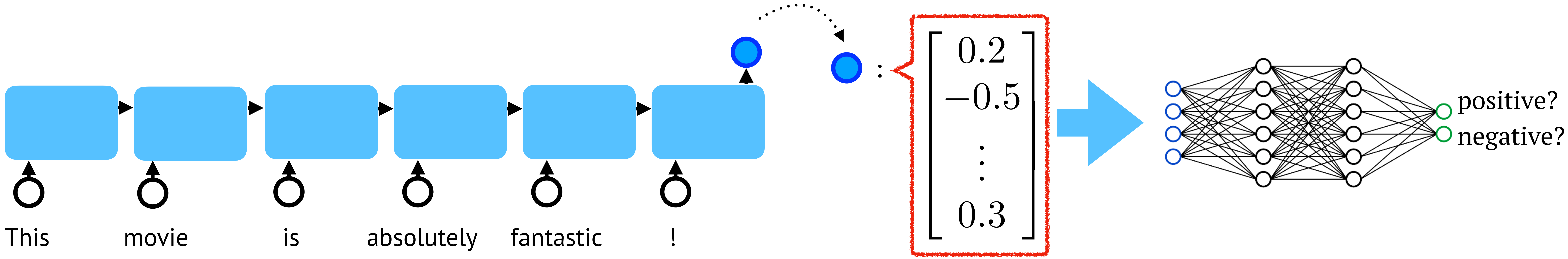
Recurrent Neural Network



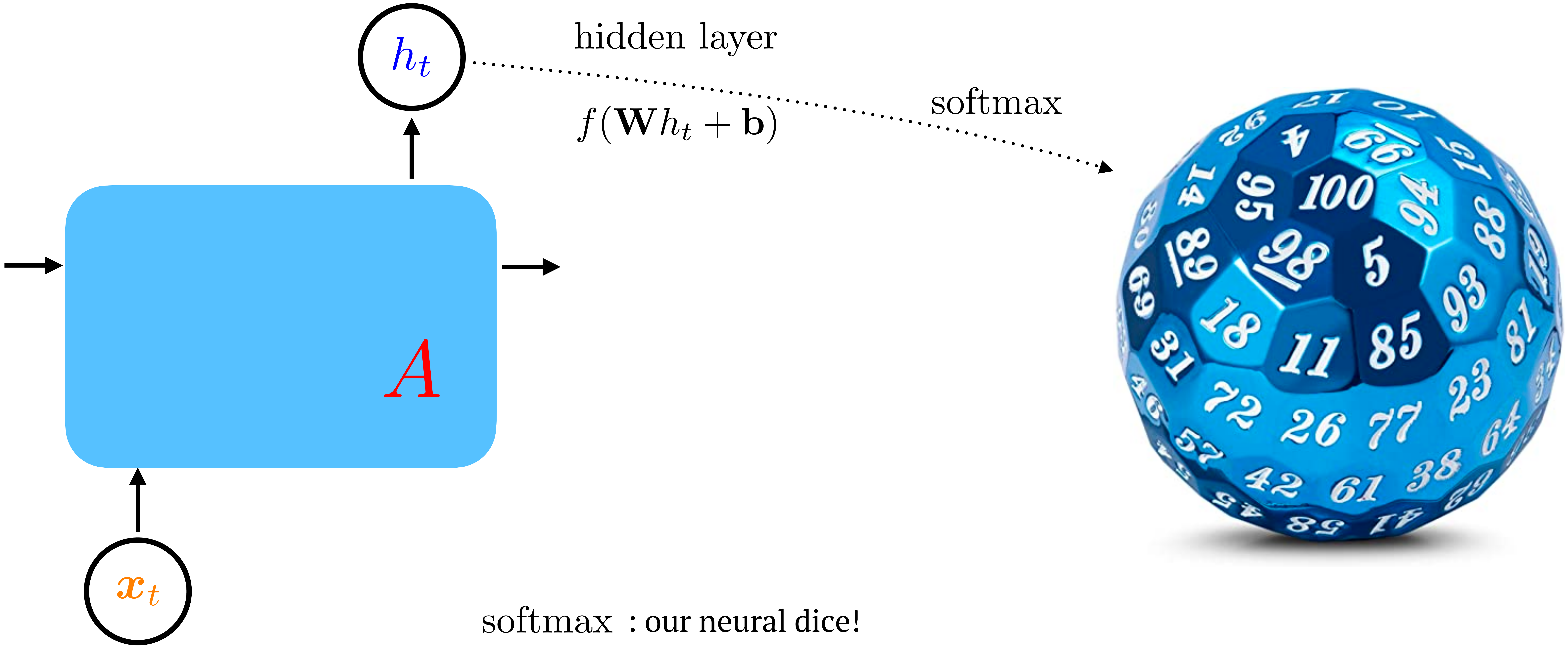
RNN as Encoder



Representation of the Sequence
(Encode into a vector representation)



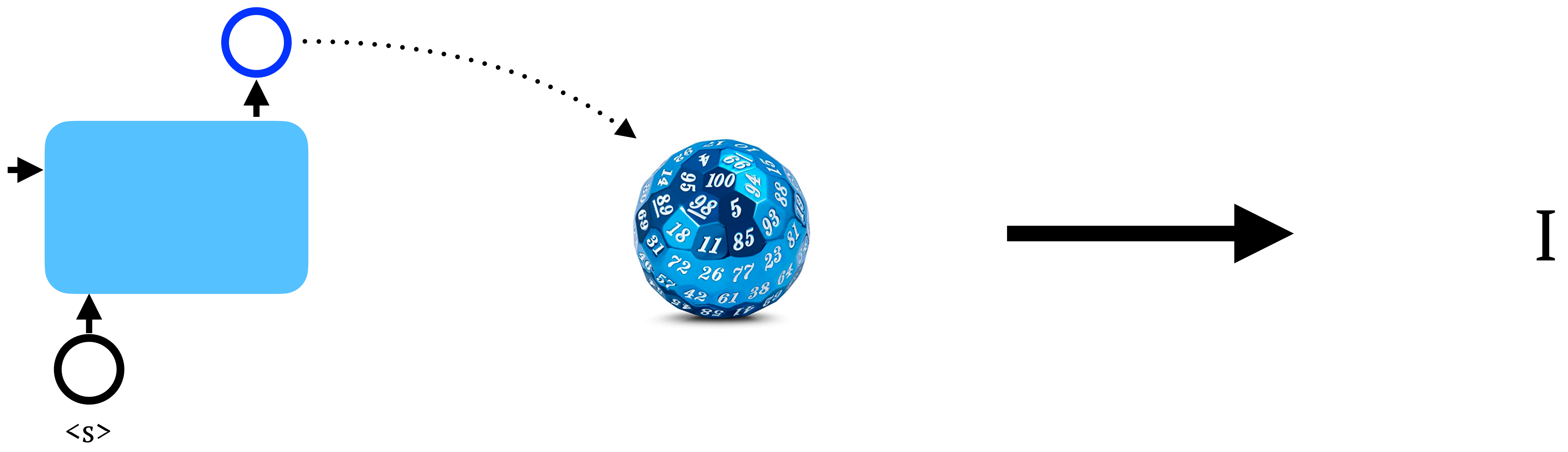
Sample a sentence from RNNLMs



softmax : our neural dice!

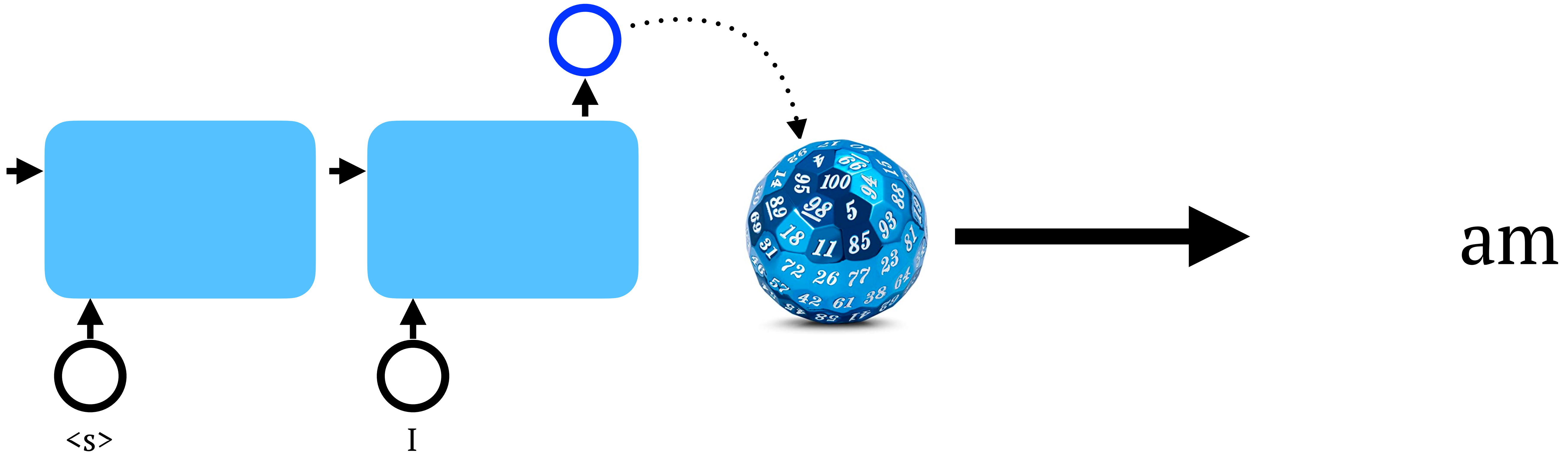
Sample a sentence from RNNLMs

I



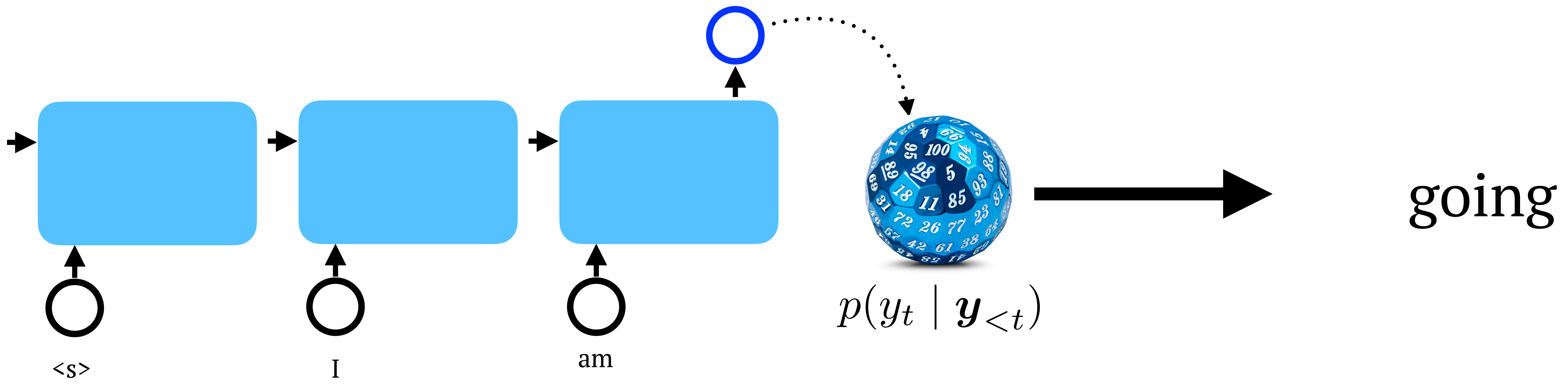
Sample a sentence from RNNLMs

I am



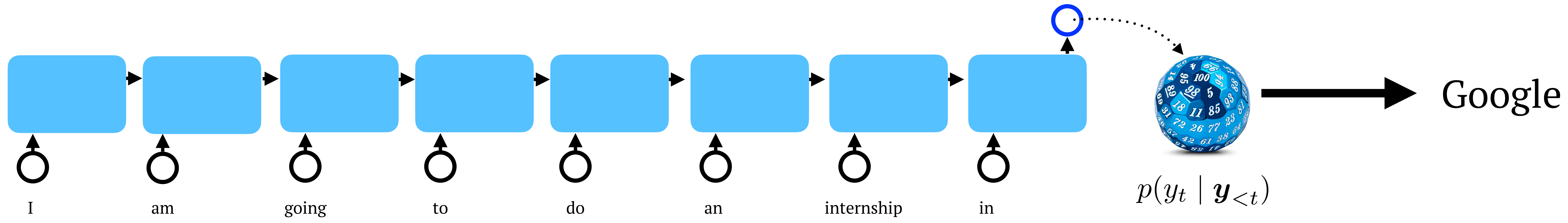
Sample a sentence from RNNLMs

I am going

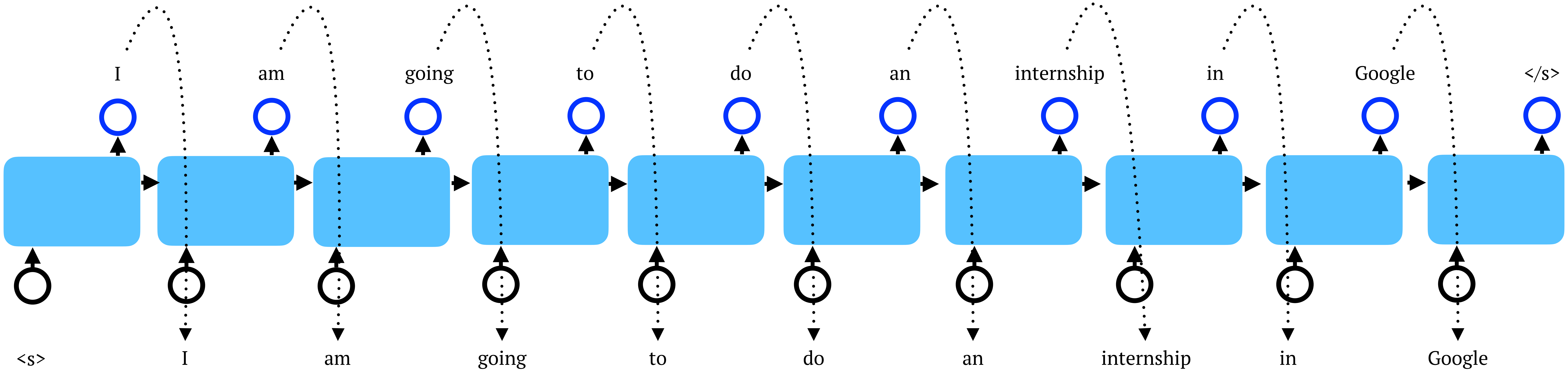


Sample a sentence from RNNLMs

I am going to do an internship in Google



RNN as Decoder (RNNLM)



$$p(y_t \mid \boldsymbol{y}_{<t})$$

Machine Translation

中秋快樂！

$\boldsymbol{x}$

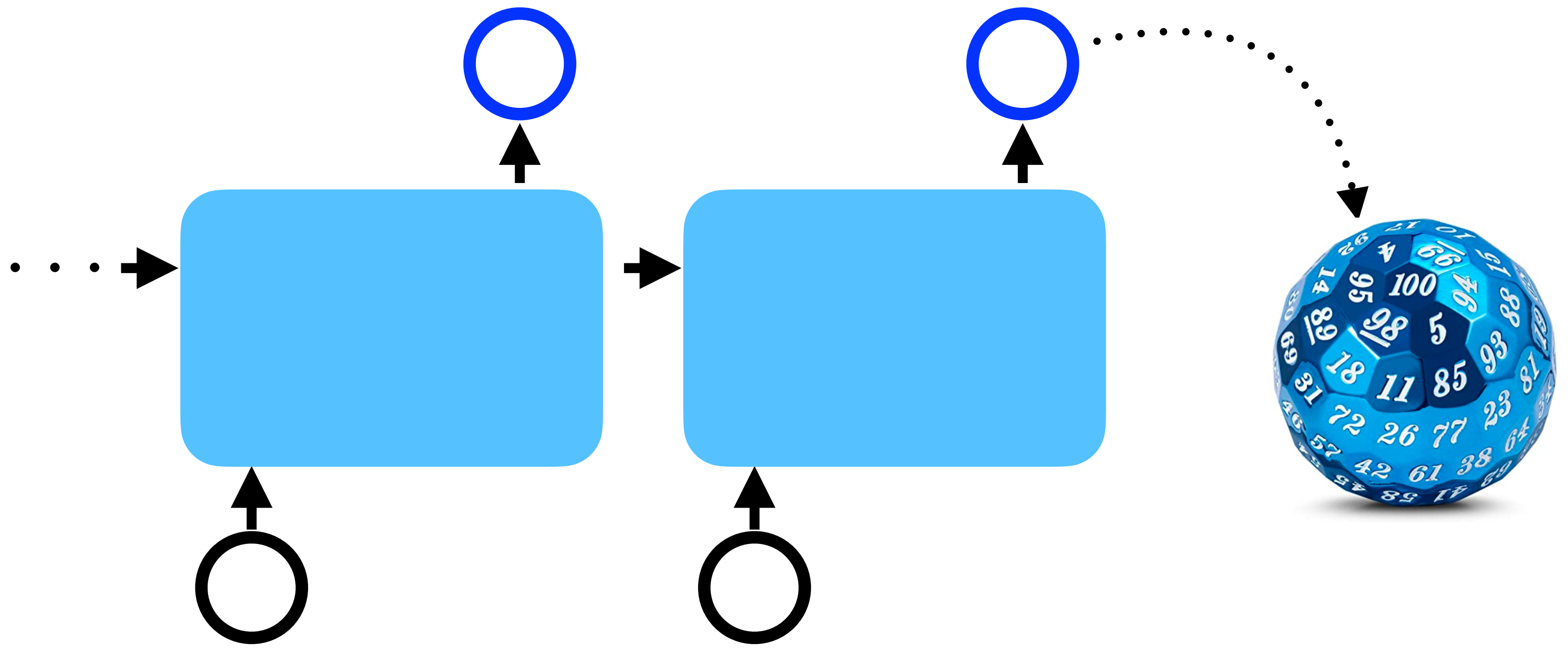
Happy mid autumn festival !

$\boldsymbol{y}$

Happy mid autumn festival !

$$p(\boldsymbol{y}) = p(y_1 \ldots y_n) = \prod_{t=1}^{n} p(y_t \mid \boldsymbol{y}_{<t})$$

$p(y_t \mid \boldsymbol{y}_{<t})$



Machine Translation

中 秋 快 樂 !　　　　Happy mid autumn festival !

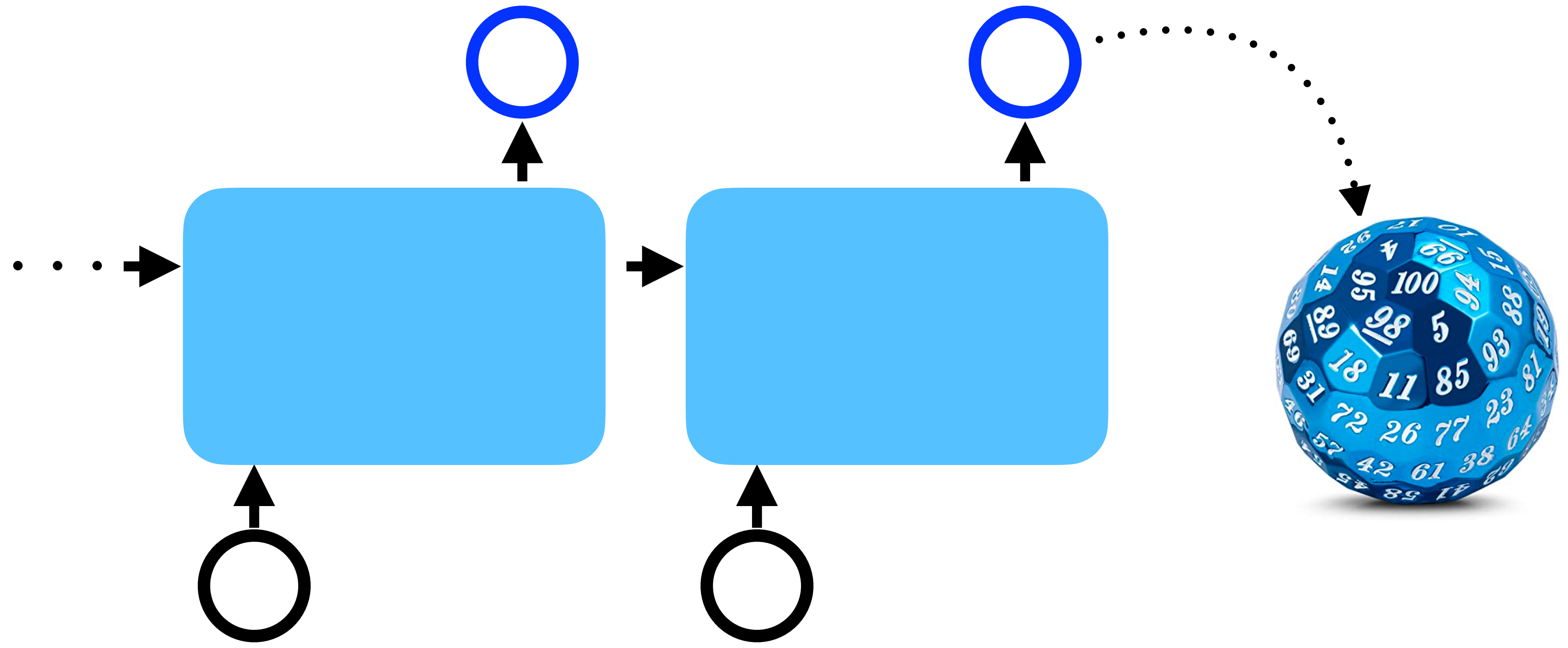
$\boldsymbol{x}$　　　　$\boldsymbol{y}$

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = p(y_1 \dots y_n \mid x_1 \dots x_m) = \prod_{t=1}^{n} p(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{x})$$

target　source

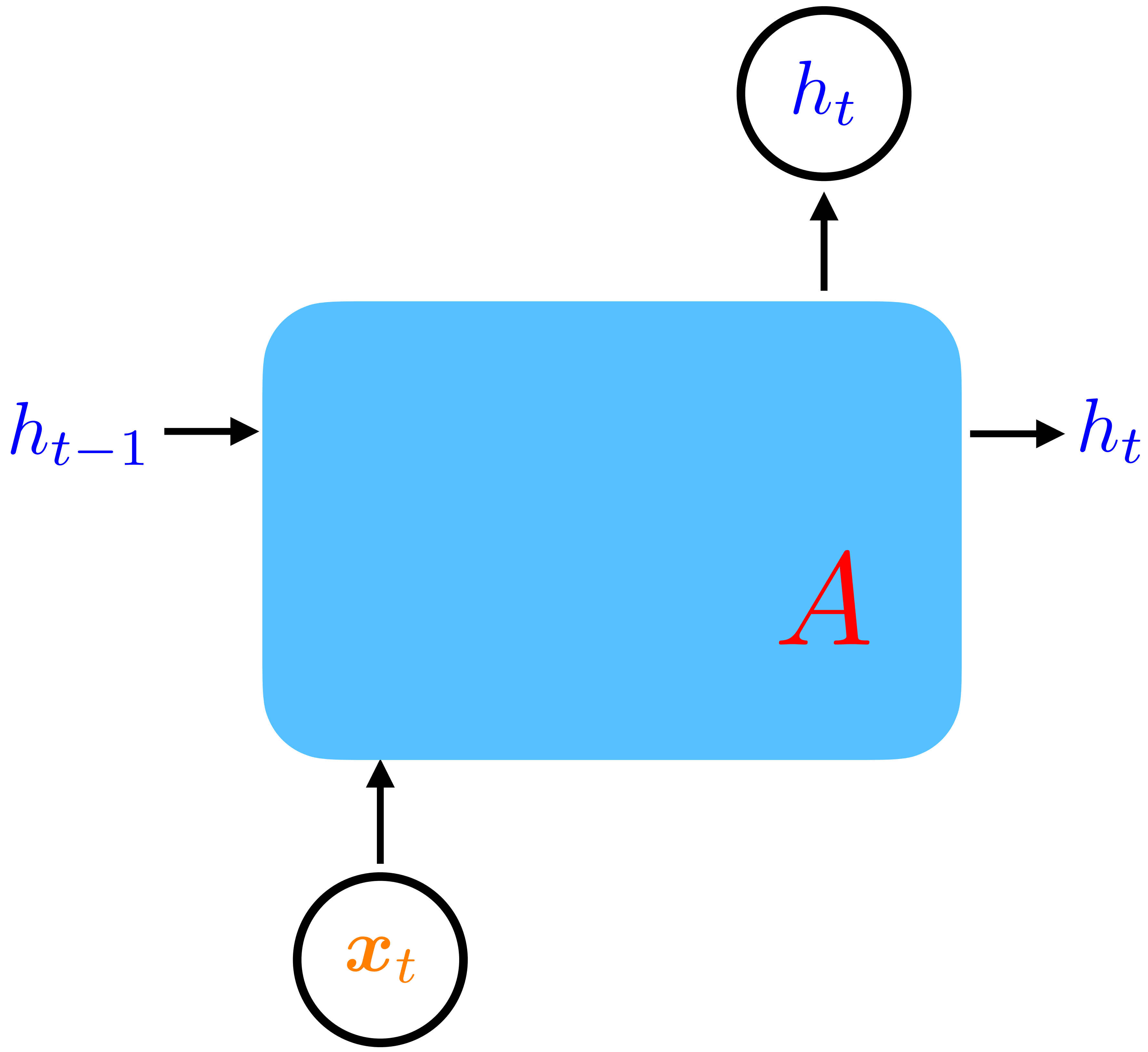
Conditional Language Model

$p(y_t \mid \boldsymbol{y}_{<t})$



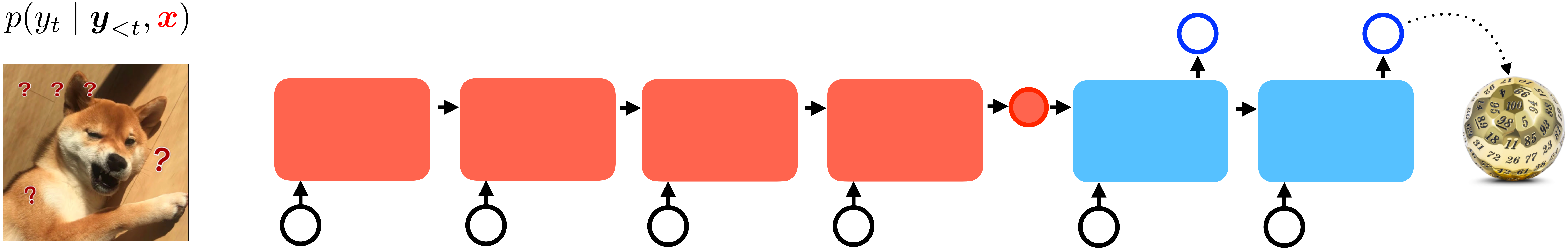
$p(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{x})$

Recurrent Neural Network



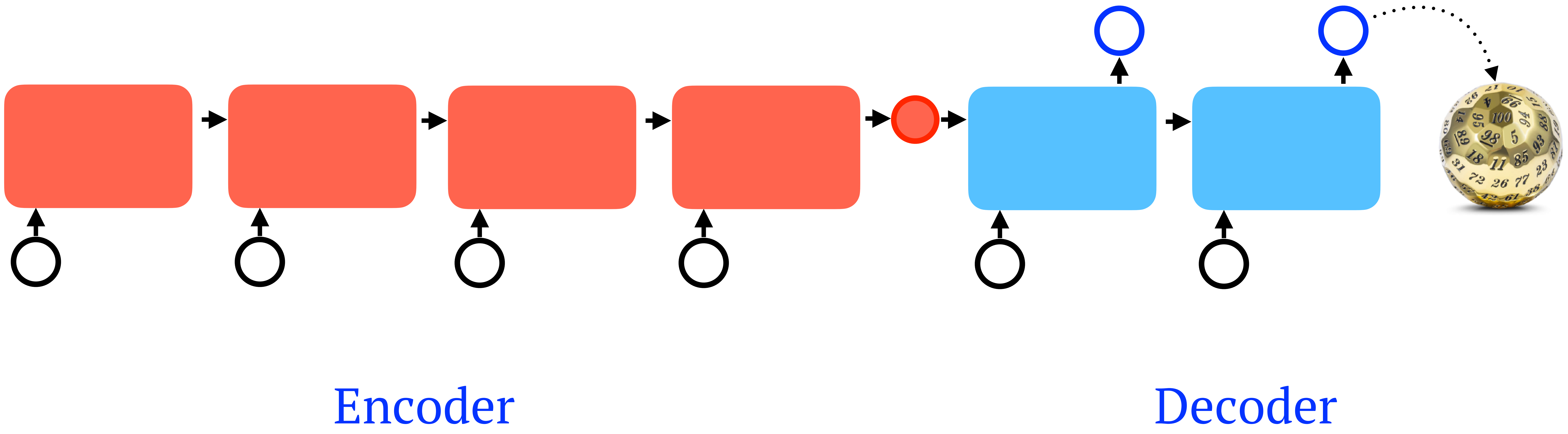
Encoder + Decoder

$p(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{x})$

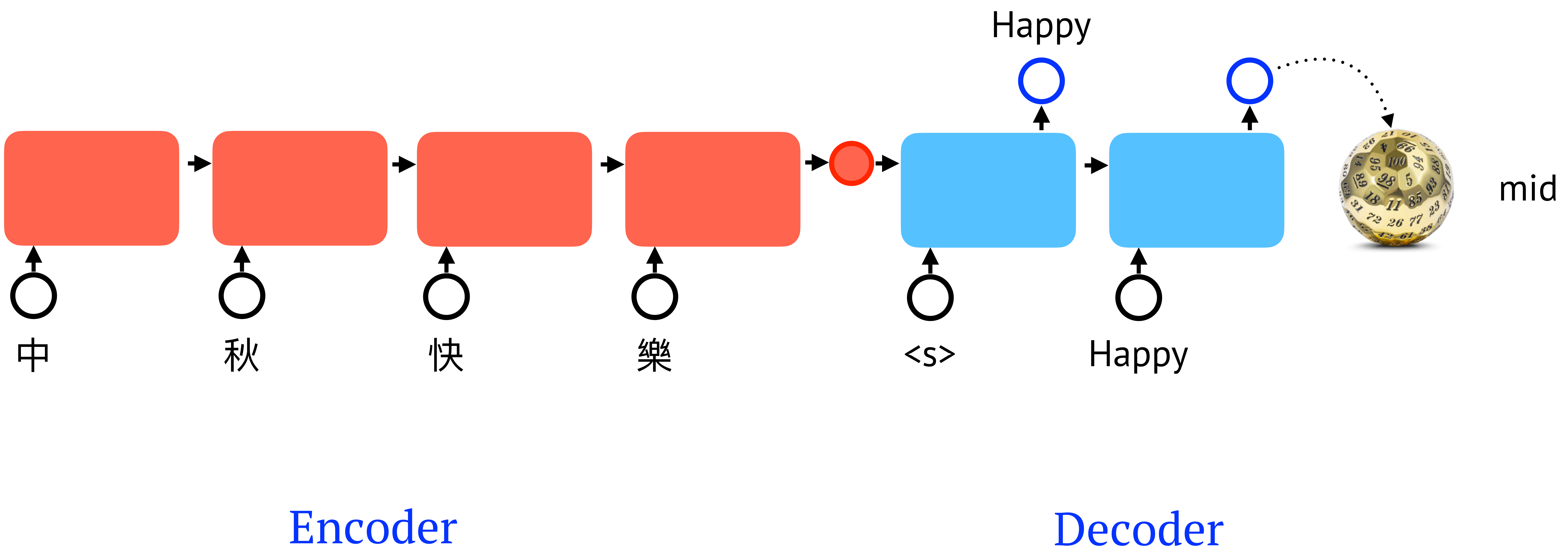


Sequence to Sequence Model

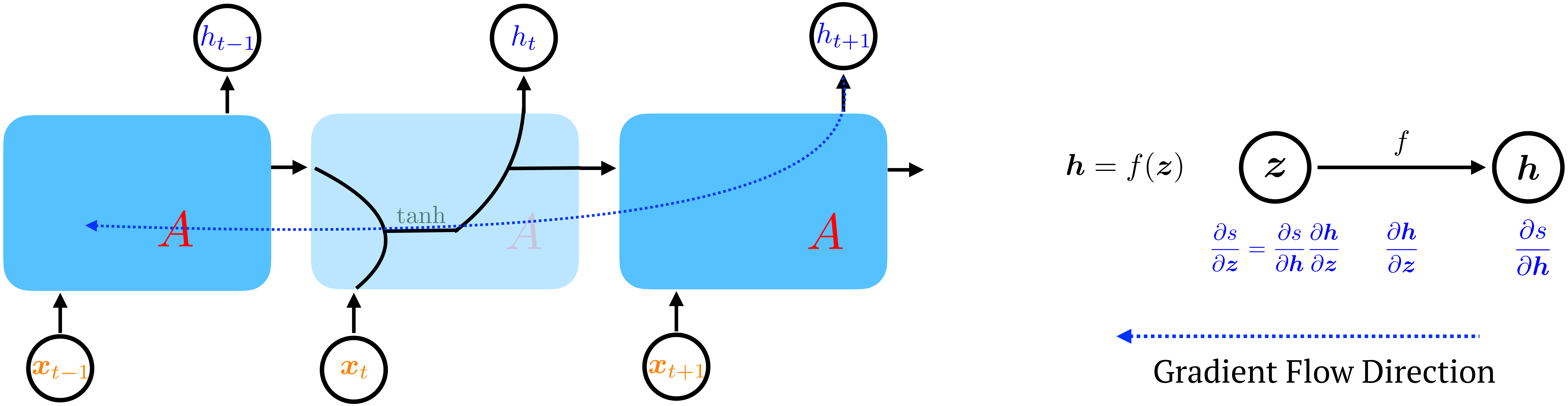
$p(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{x})$



Sequence to Sequence Model

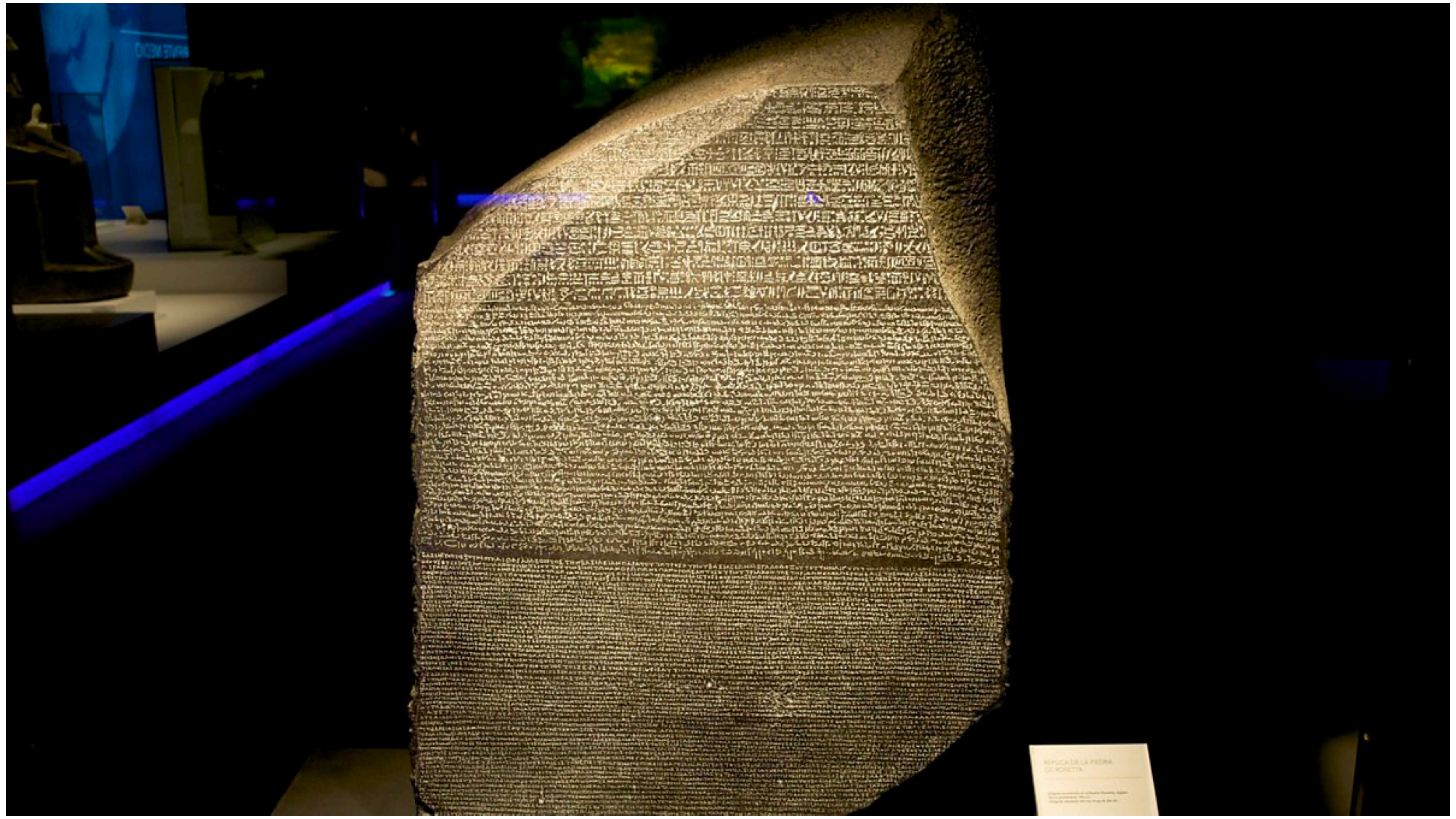


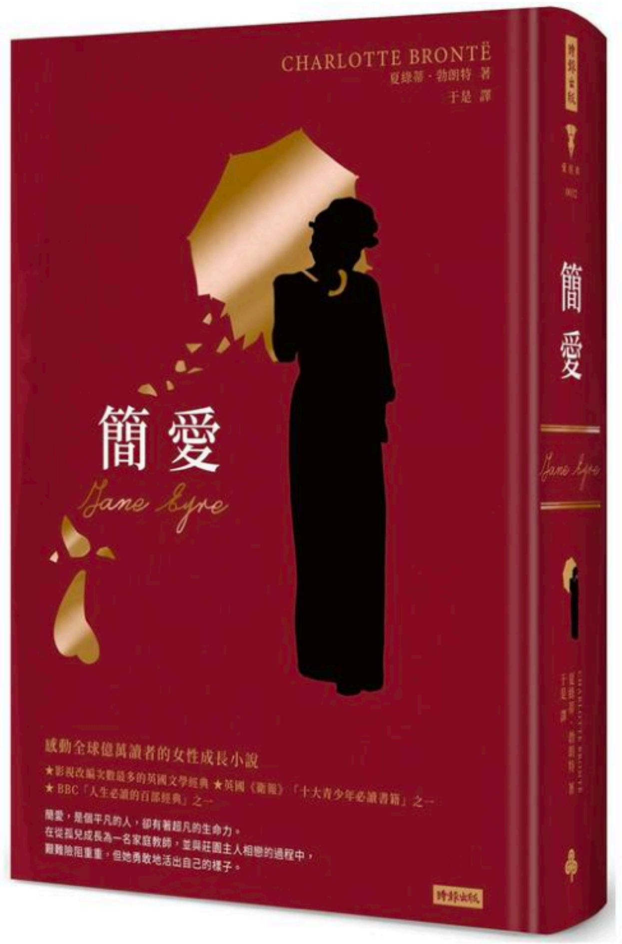
Vanishing Gradient in RNNs

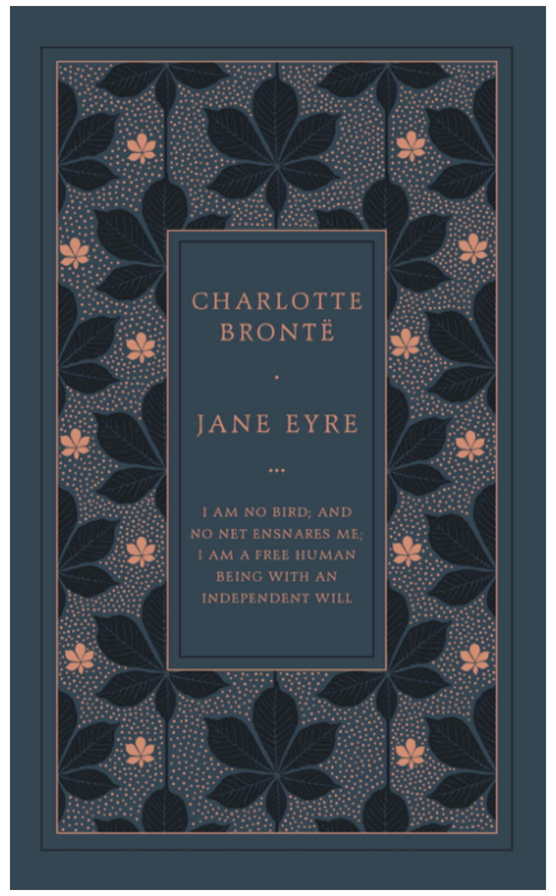


In general, the longer the path, the smaller the gradient signal.

Alignment in Machine Translation





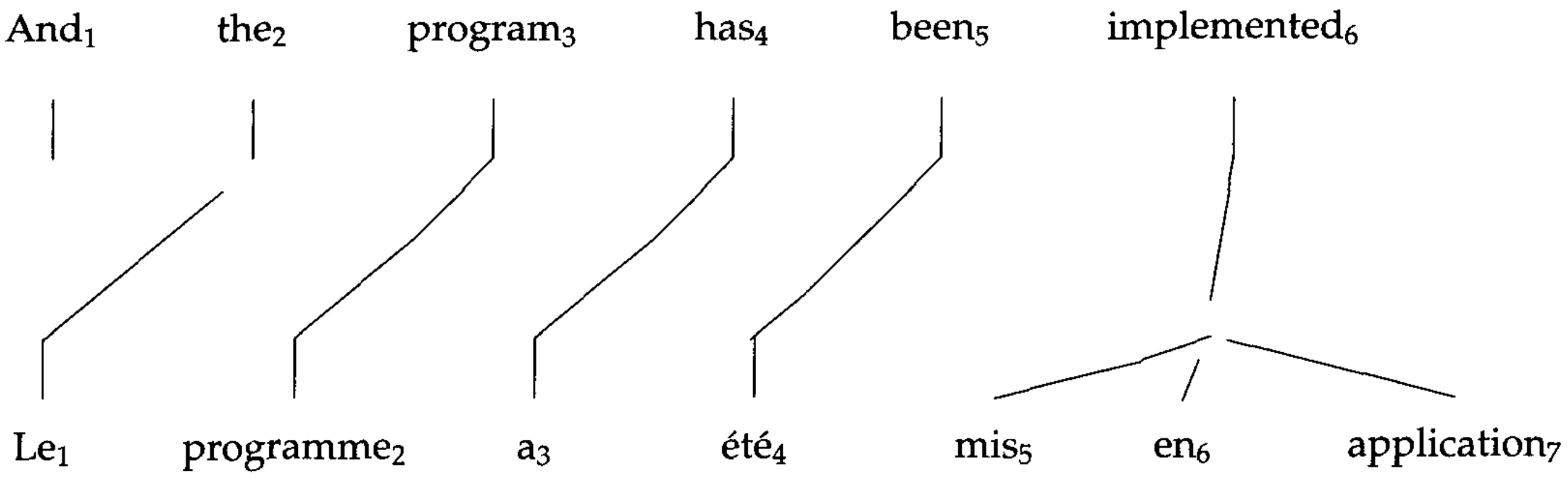


Parallel Corpus

$$p(\boldsymbol{y} \mid \boldsymbol{x})$$

target source

Alignment in Machine Translation

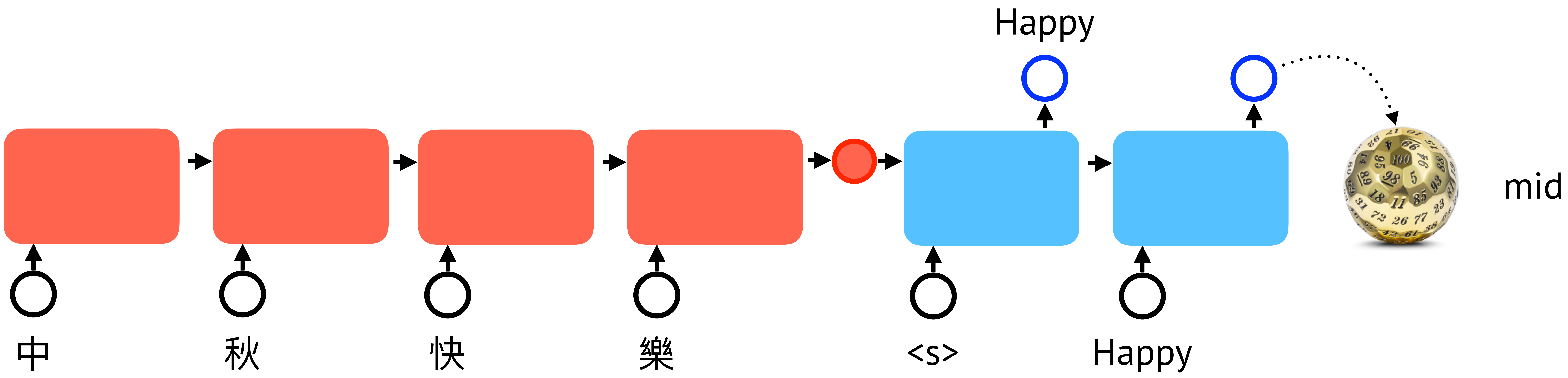


Some words might have no “counter-part”.

Alignment can be many-to-one (or one-to-many).

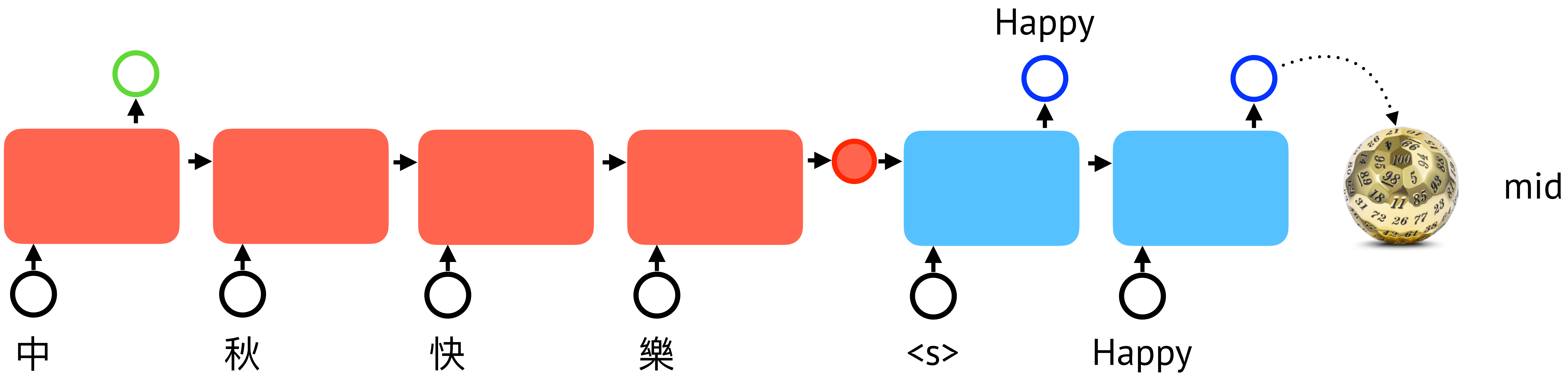
(Brown et al, 1993)

Sequence to Sequence Model

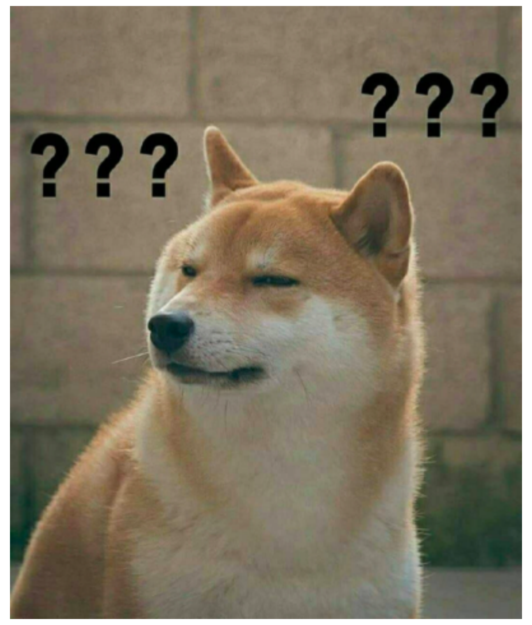


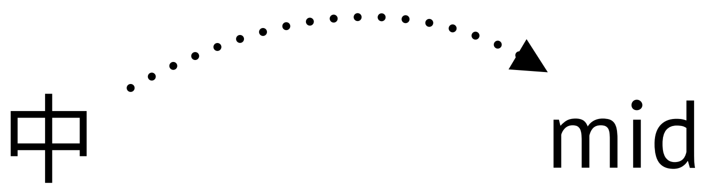
$$p(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{x})$$

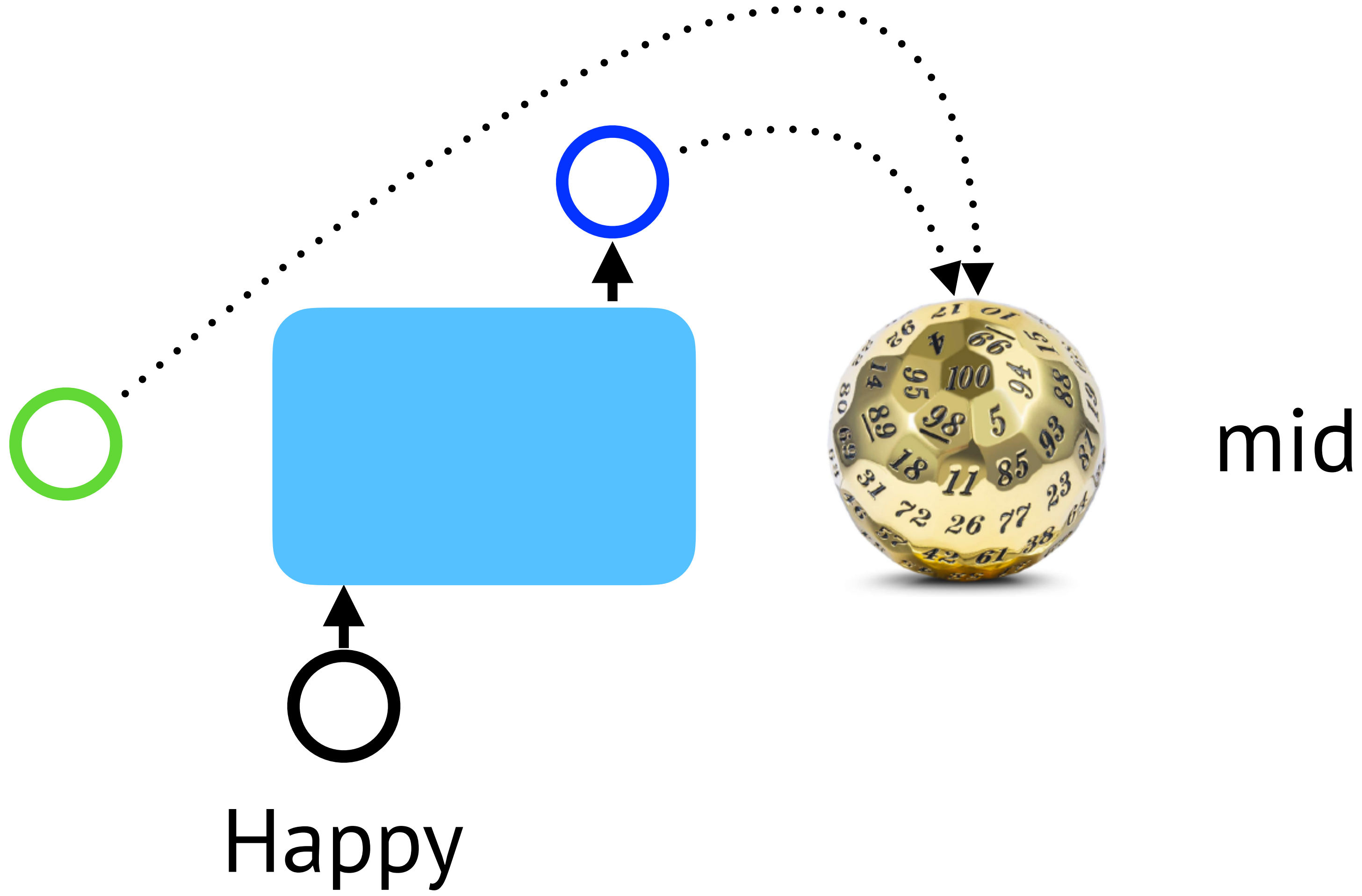
Sequence to Sequence Model



$p(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{x})$

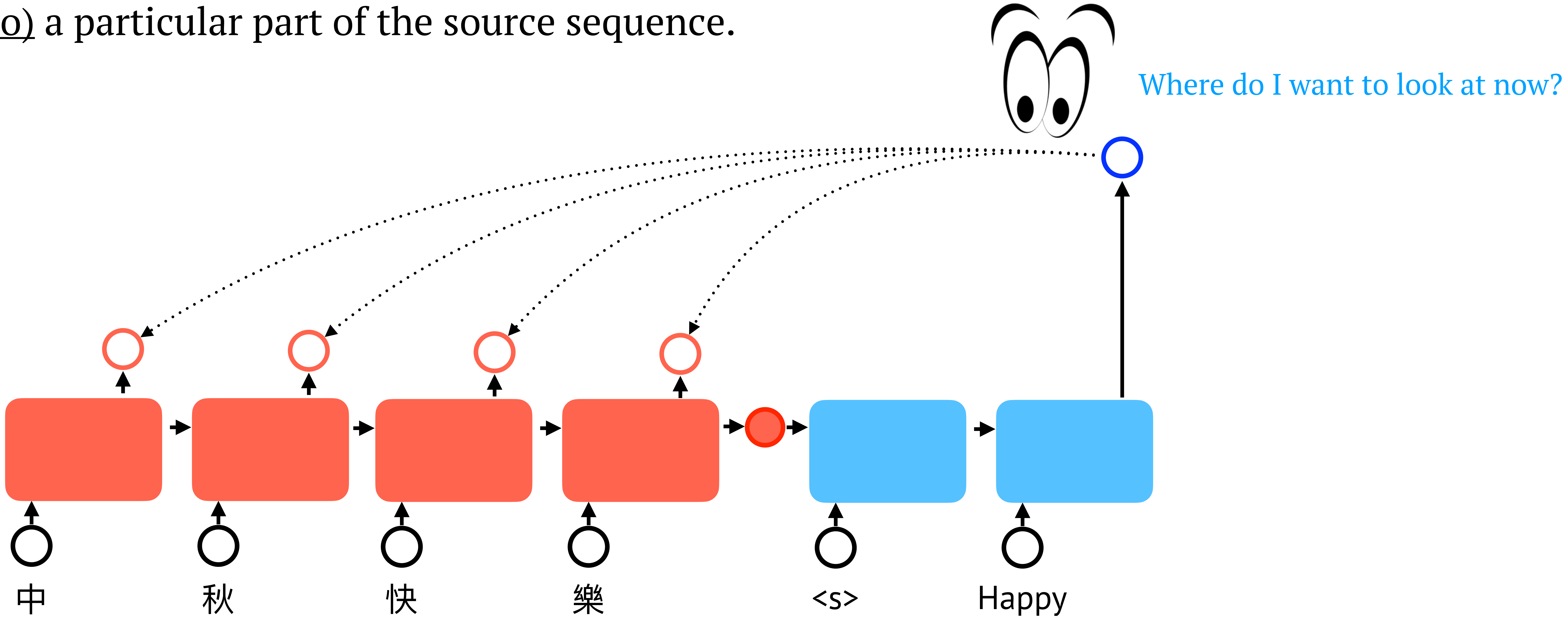






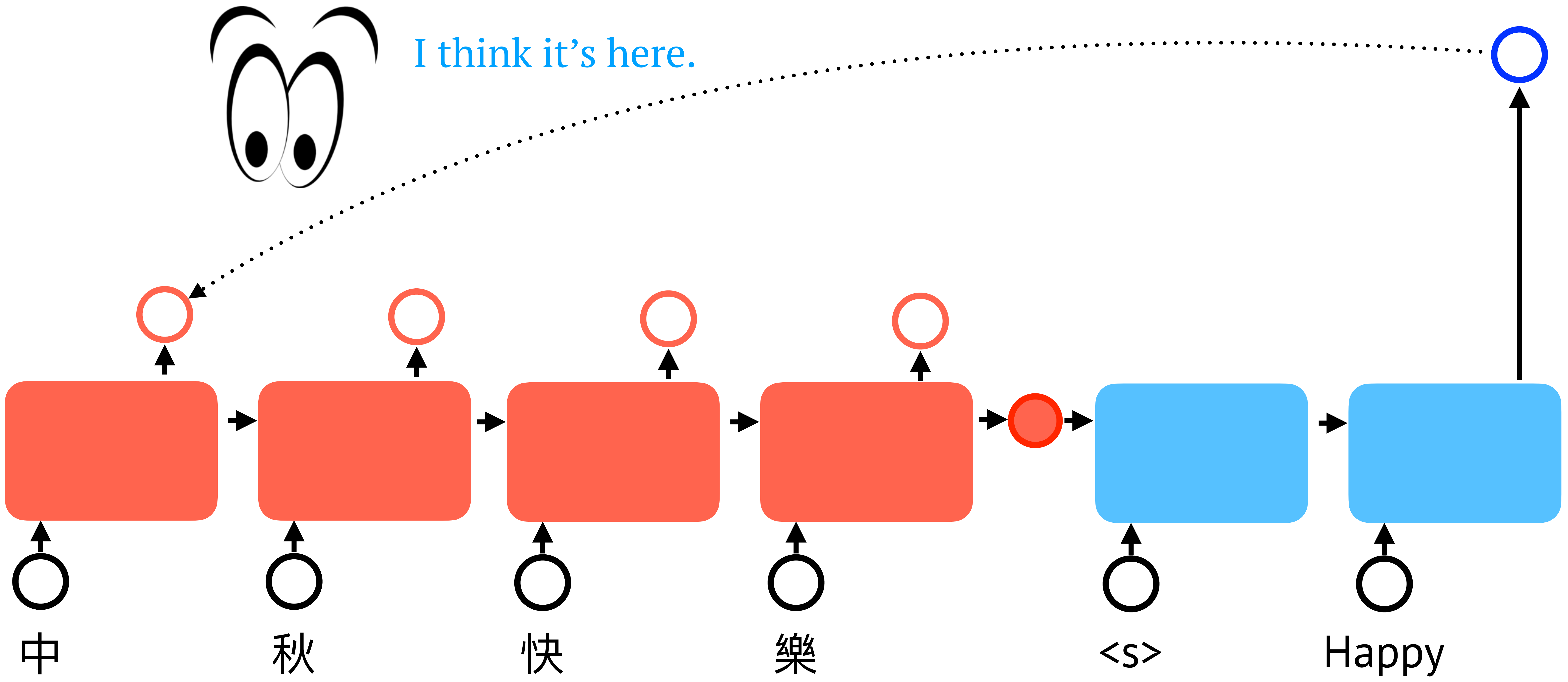
Attention Mechanism

Use direct connection to the encoder to focus on (attend to) a particular part of the source sequence.



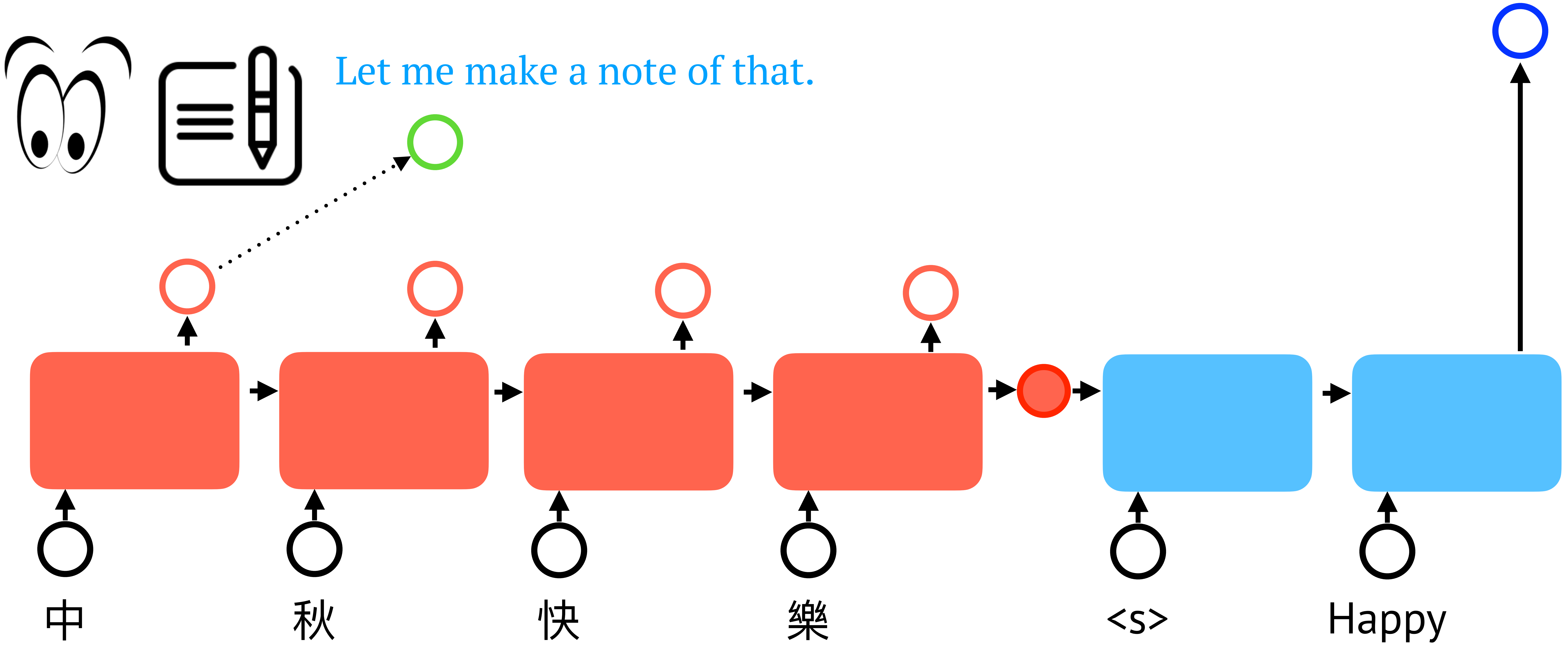
Attention Mechanism

Use direct connection to the encoder to focus on (attend to) a particular part of the source sequence.



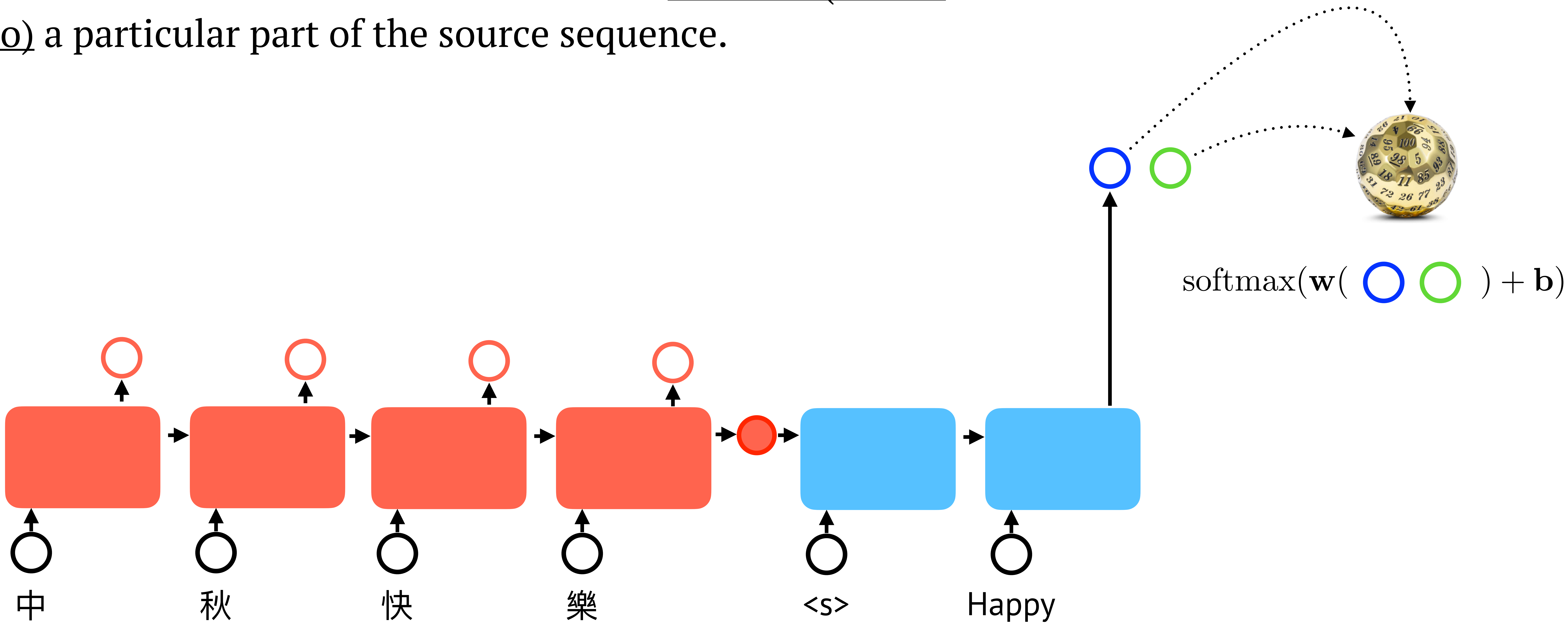
Attention Mechanism

Use direct connection to the encoder to focus on (attend to) a particular part of the source sequence.

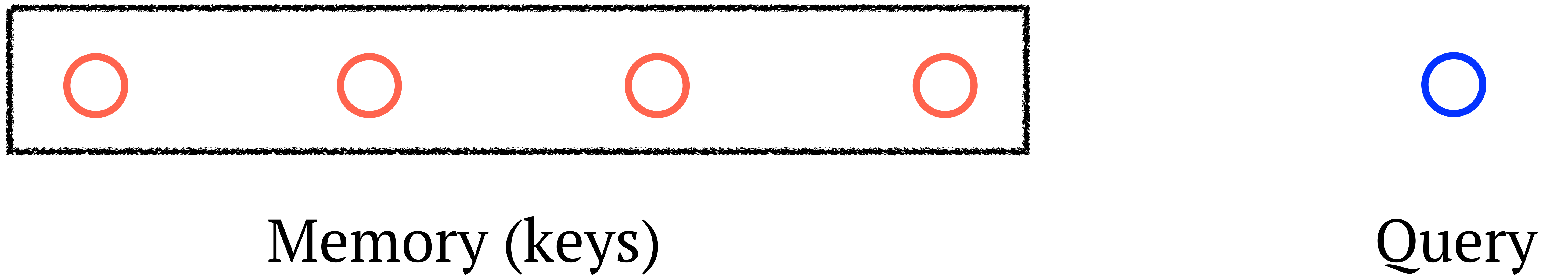


Attention Mechanism

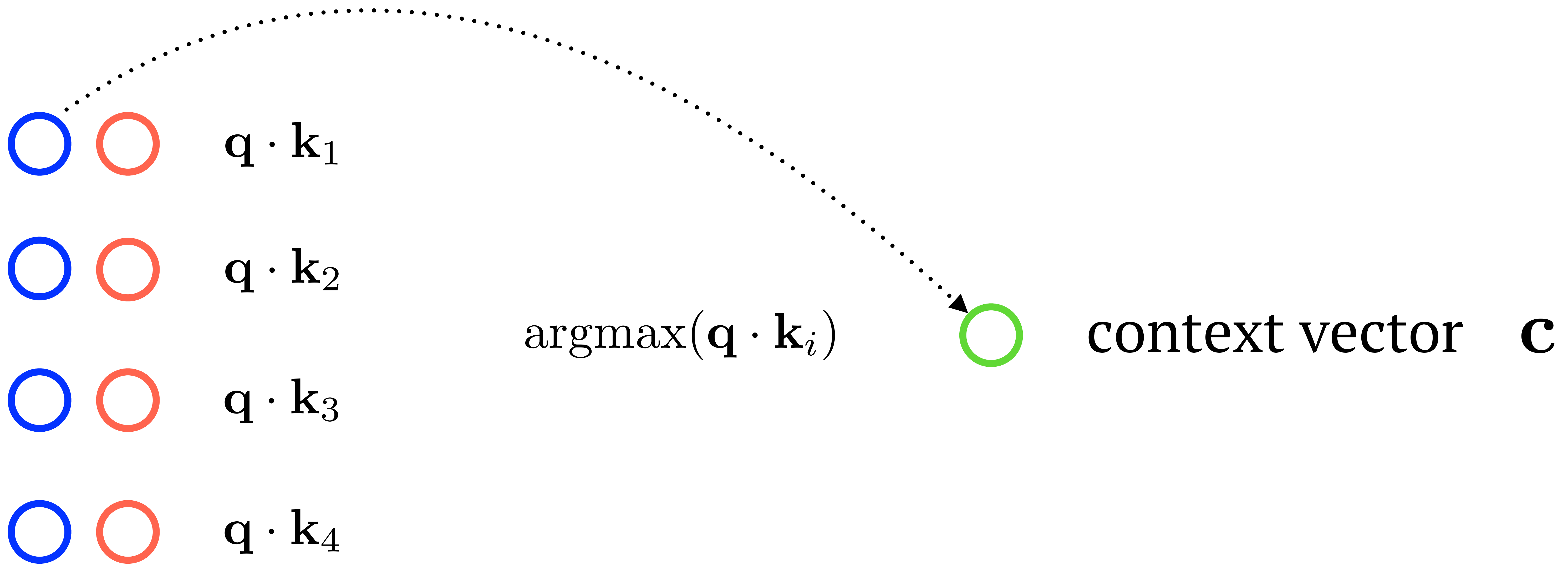
Use direct connection to the encoder to <u>focus on (attend to)</u> a particular part of the source sequence.



Memory Abstraction



Task: Finding the most "relevant" item in the memory.

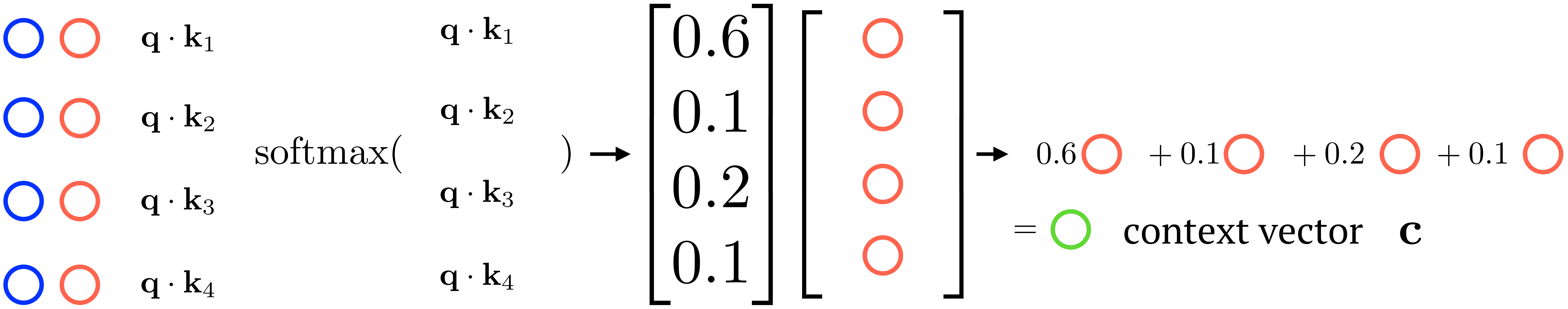


Dot-Product-Softmax Attention

Memory (keys)

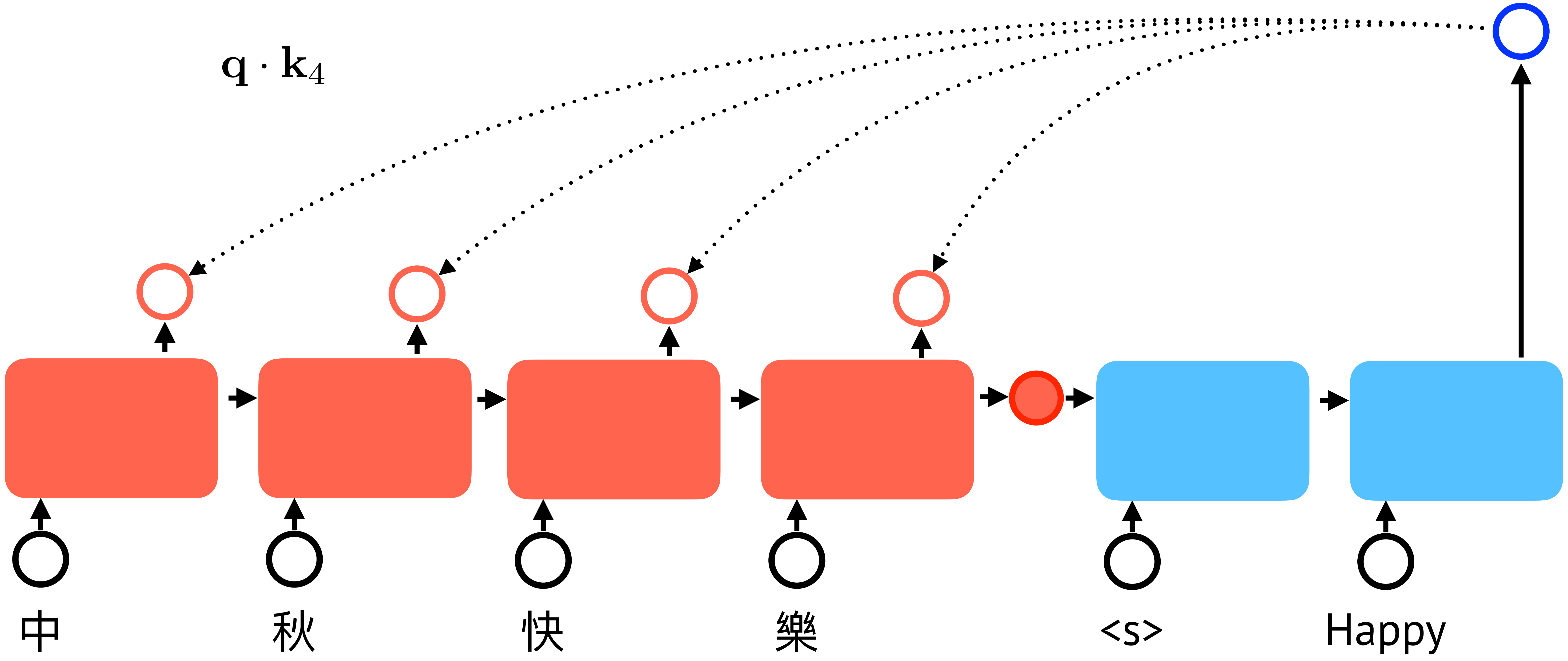
Query

Task: Finding the most “relevant” item in the memory.



Attention Mechanism

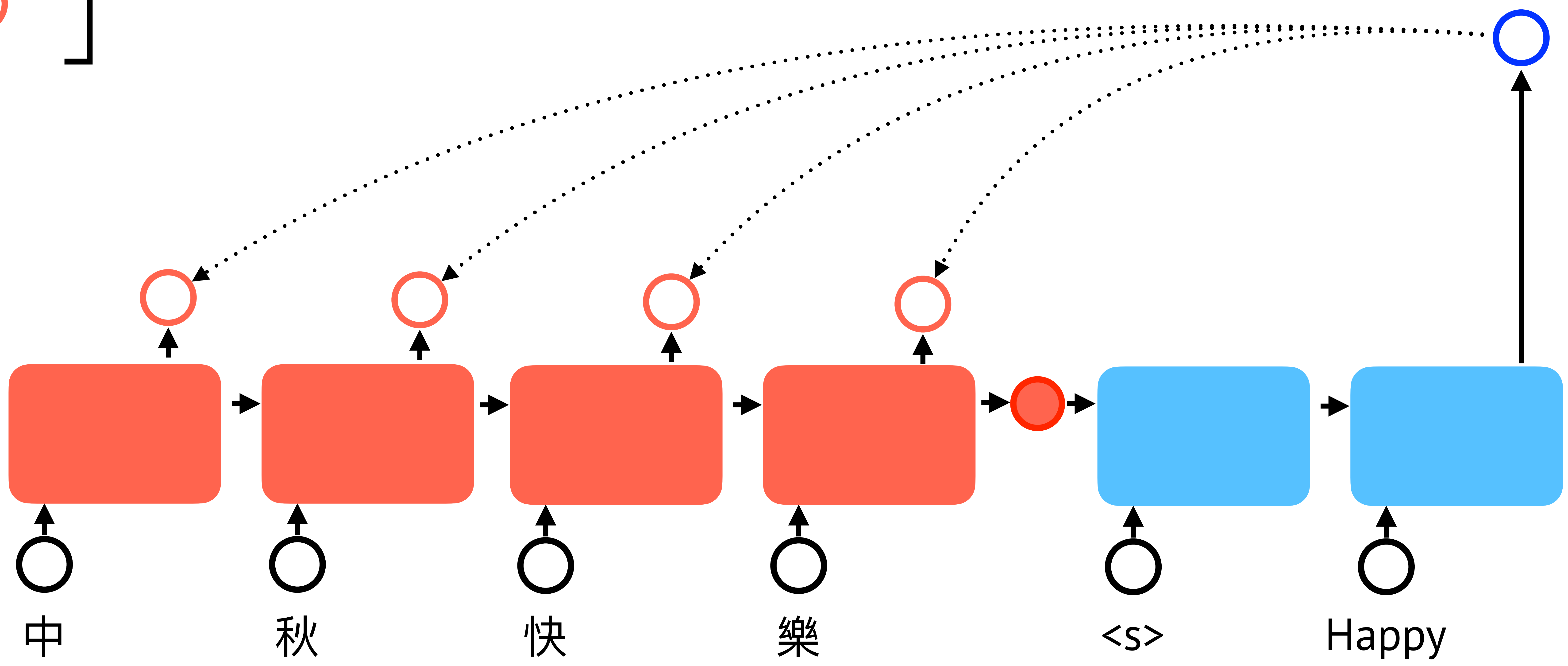
$$\mathrm{softmax}\begin{pmatrix} \mathbf{q}\cdot\mathbf{k}_1 \\ \mathbf{q}\cdot\mathbf{k}_2 \\ \mathbf{q}\cdot\mathbf{k}_3 \\ \mathbf{q}\cdot\mathbf{k}_4 \end{pmatrix}$$



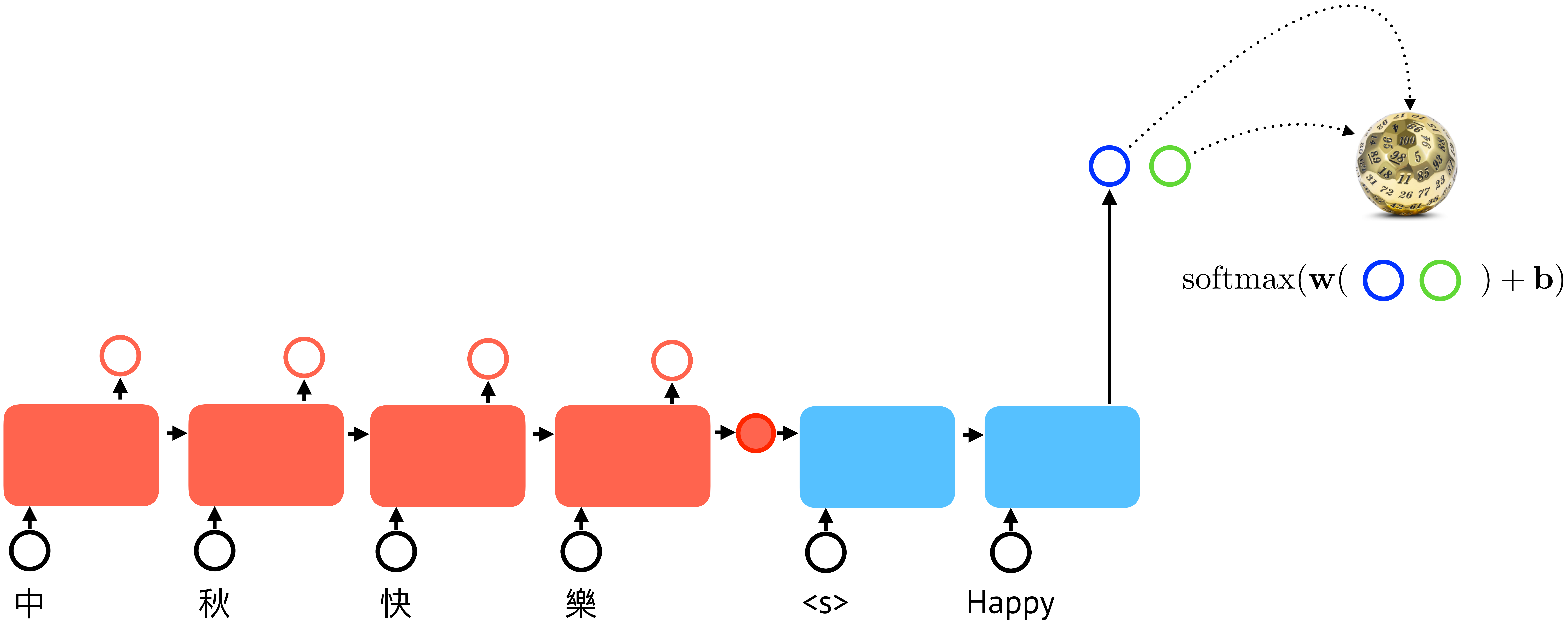
Attention Mechanism

$$\begin{bmatrix}0.6\\0.1\\0.2\\0.1\end{bmatrix}\begin{bmatrix}\bigcirc\\\bigcirc\\\bigcirc\\\bigcirc\end{bmatrix} \rightarrow 0.6\bigcirc + 0.1\bigcirc + 0.2\bigcirc + 0.1\bigcirc$$

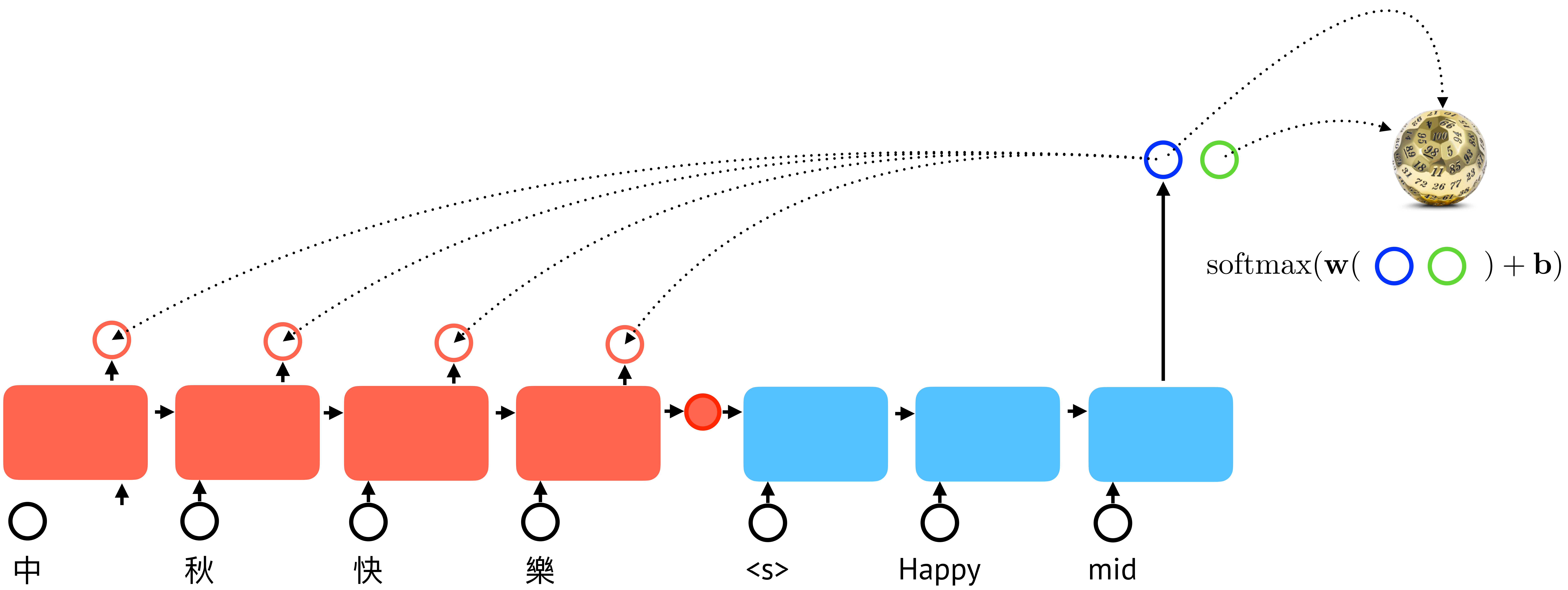
$= \bigcirc$ context vector $\mathbf{c}$



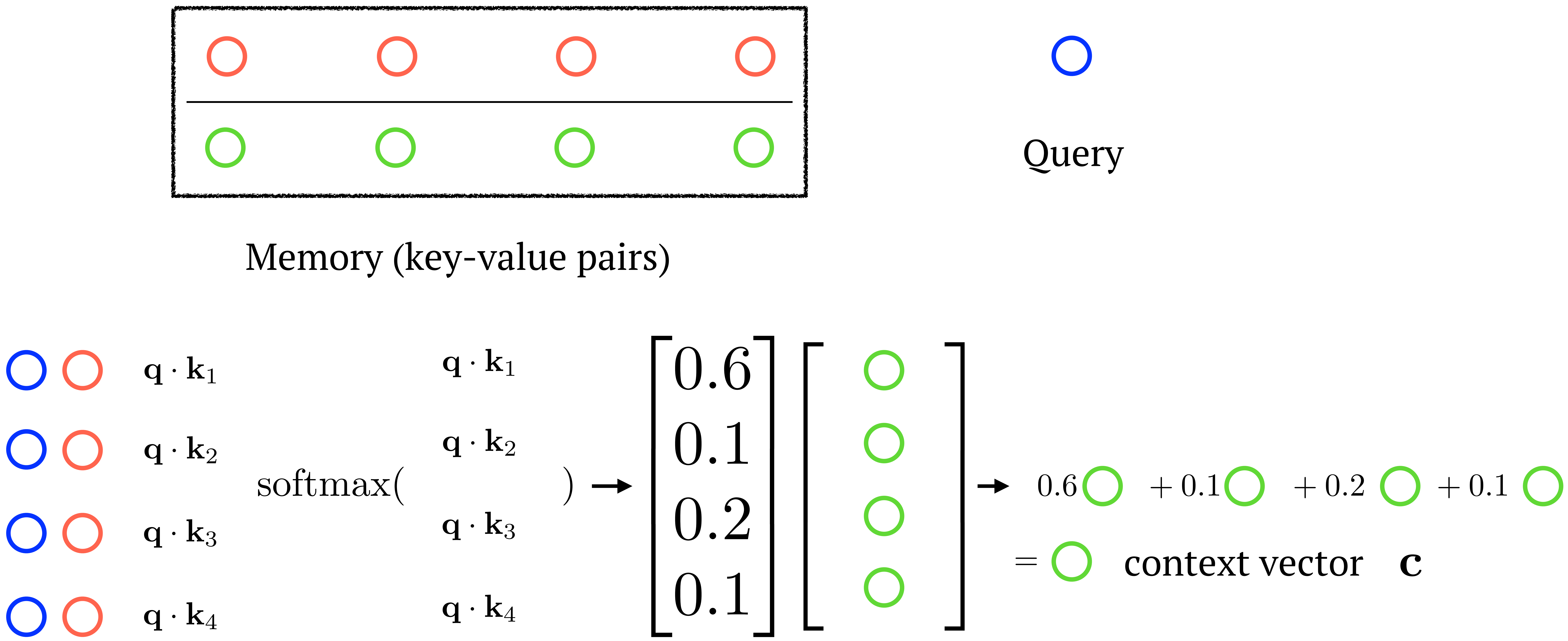
Attention Mechanism



Attention Mechanism



Dot-Product-Softmax Attention



Dot-Product-Softmax Attention

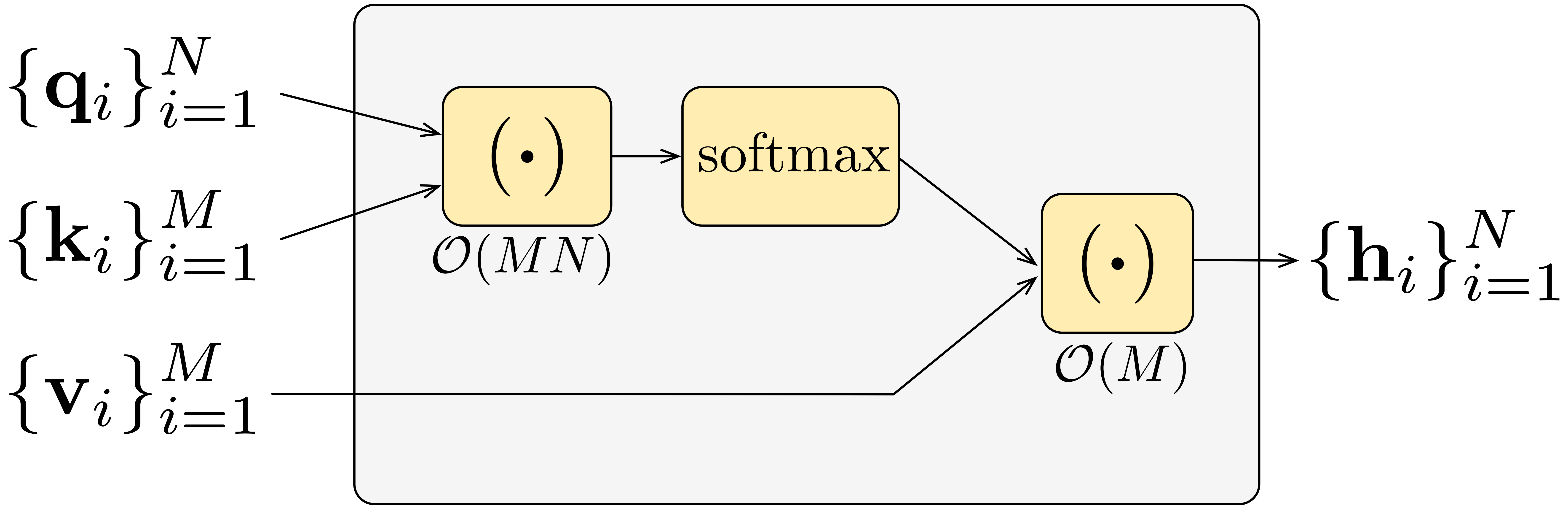
similarity

$$\underbrace{\sum_{m=1}^{M} \underbrace{\frac{\boxed{\exp\left(\boldsymbol{q}_n \boldsymbol{k}_m\right)}}{\sum_{m'=1}^{M} \exp\left(\boldsymbol{q}_n \boldsymbol{k}_{m'}\right)}}_{\text{normalized similarity}} \boldsymbol{v}_m^\top}_{\text{weighted sum}} = \mathbf{V}^\top \mathrm{softmax}(\mathbf{K}\boldsymbol{q}_n)$$

normalized similarity

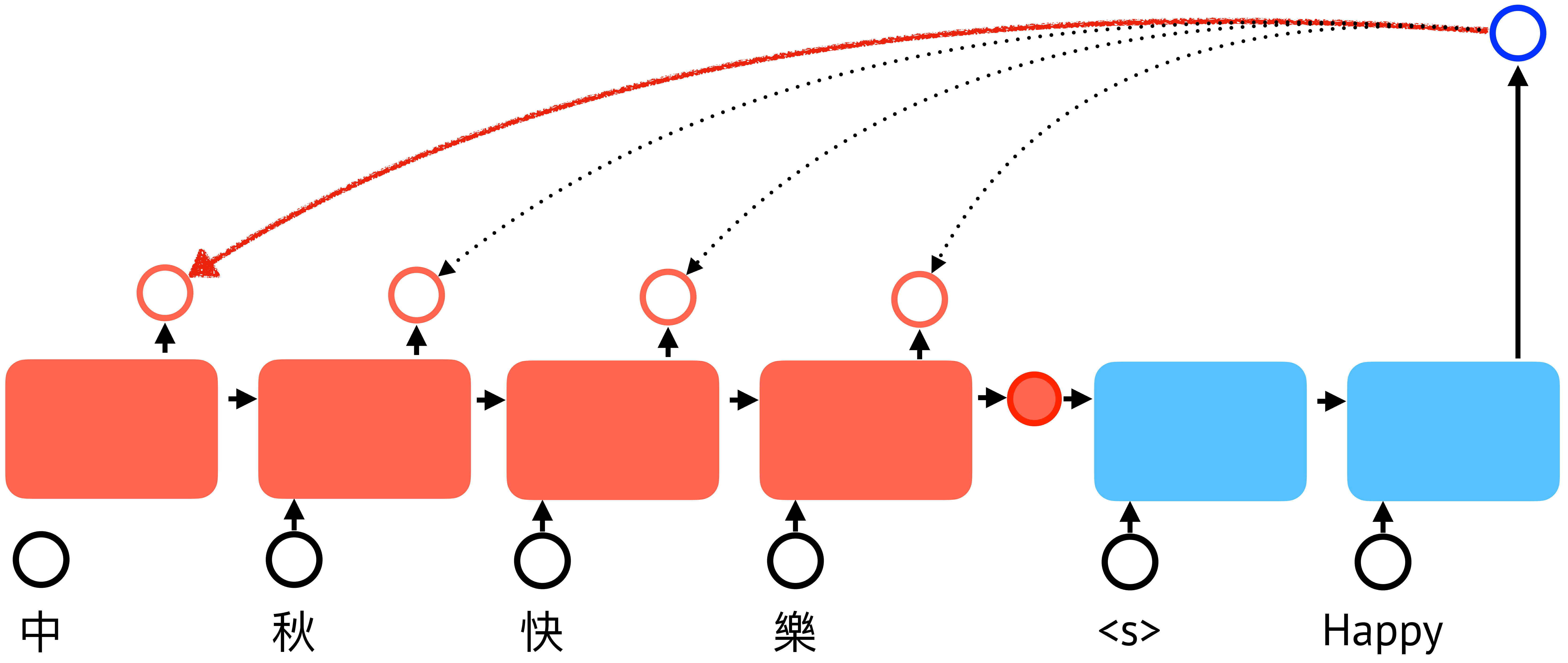
weighted sum

Computational Complexity



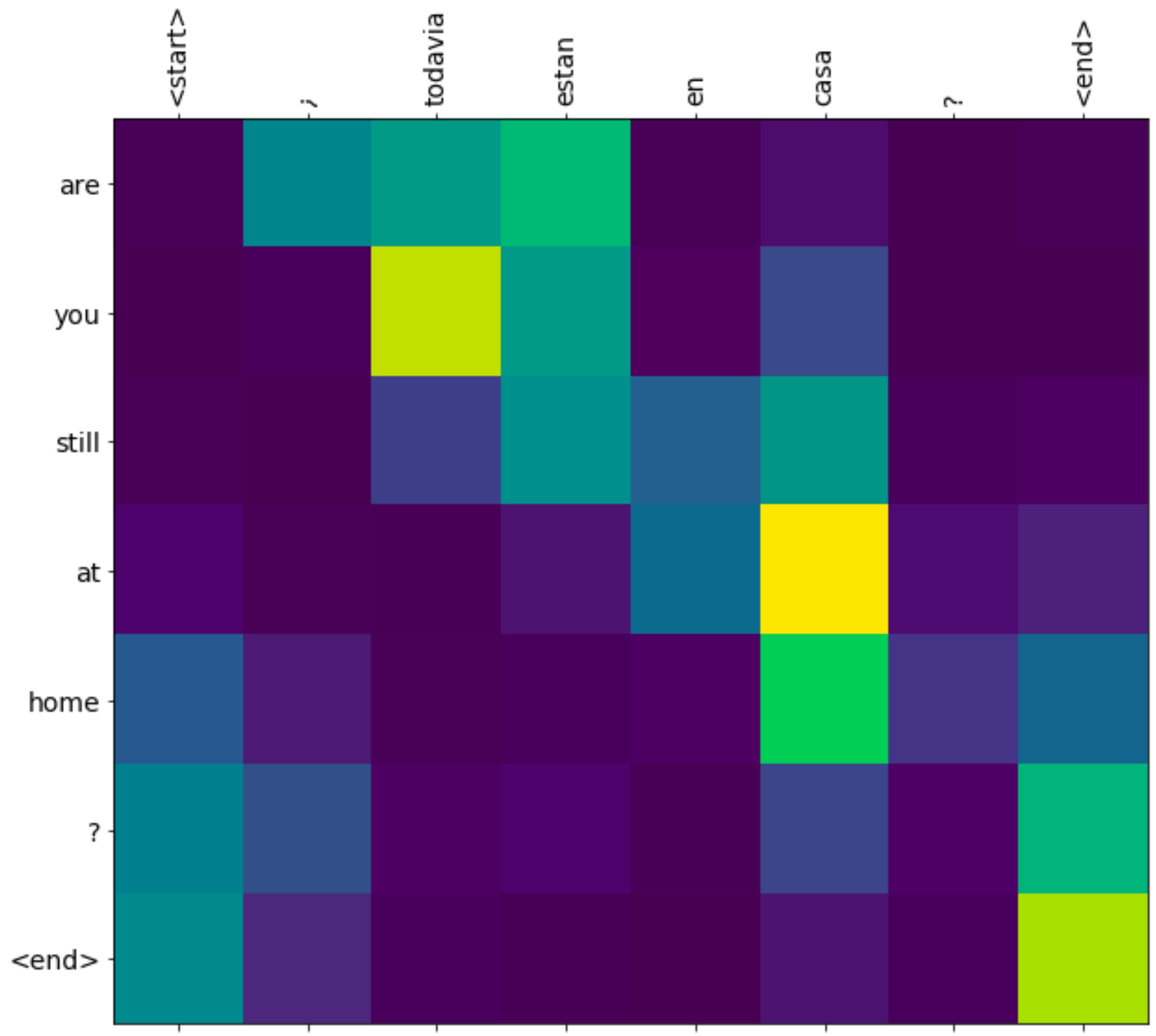
Attention Mechanism

It helps with vanishing gradient problem.



Attention Mechanism

It offers some interpretability.



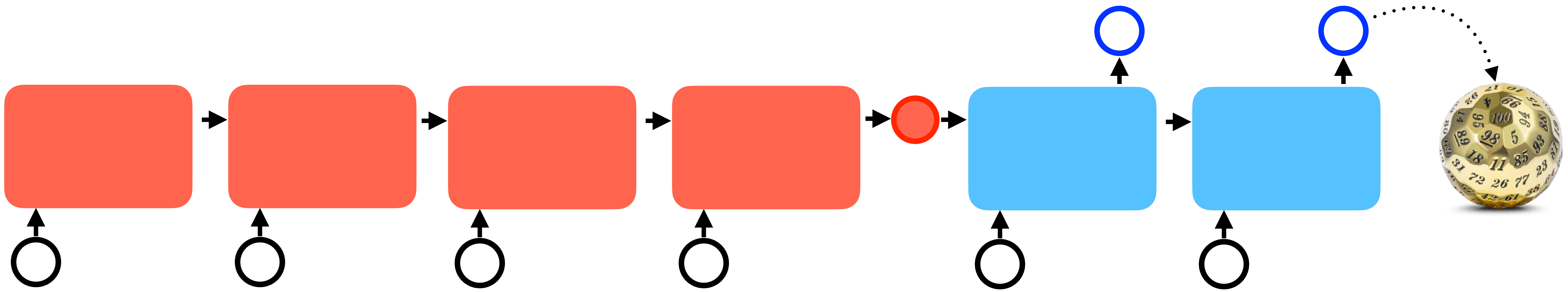
https://www.tensorflow.org/text/tutorials/nmt_with_attention

Generation as (Conditional) Language Modeling

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = p(y_1 \ldots y_n \mid x_1 \ldots x_m) = \prod_{t=1}^{n} p(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{x})$$

target source

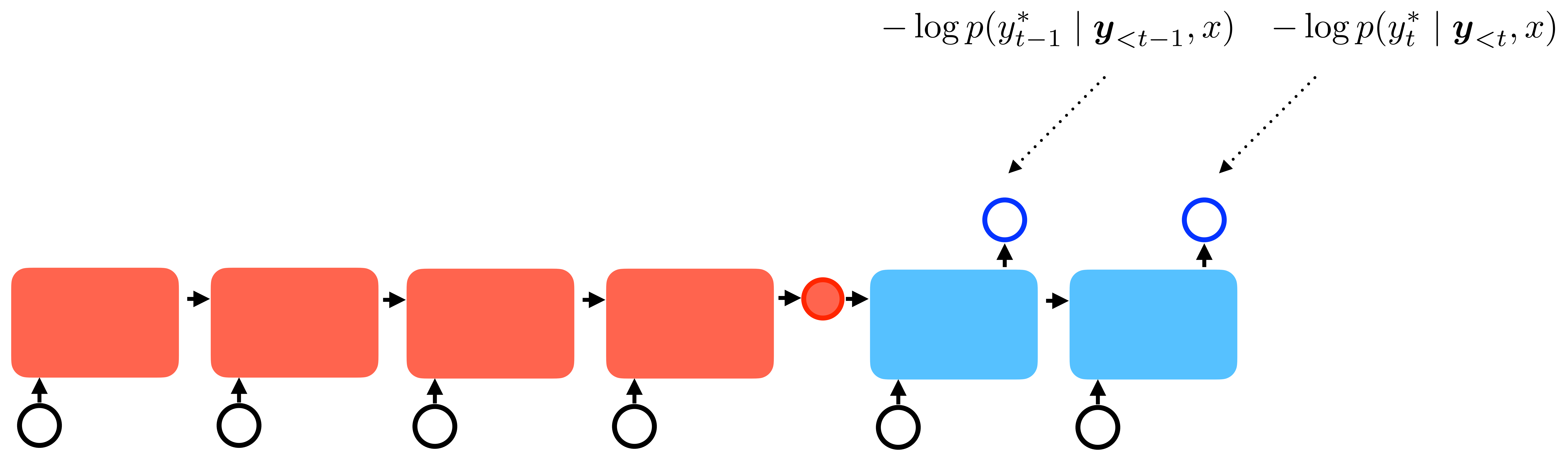
Conditional Language Model



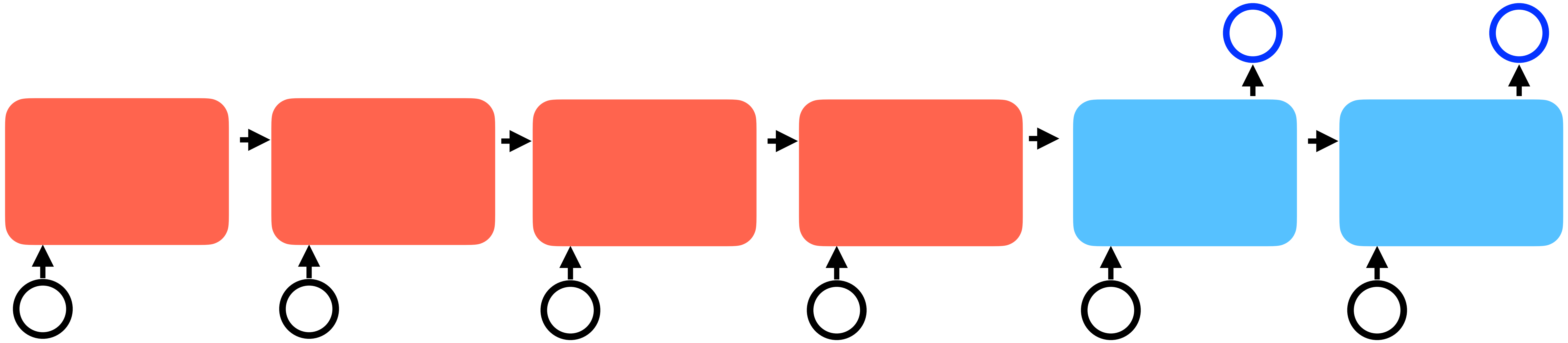
$p(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{x})$

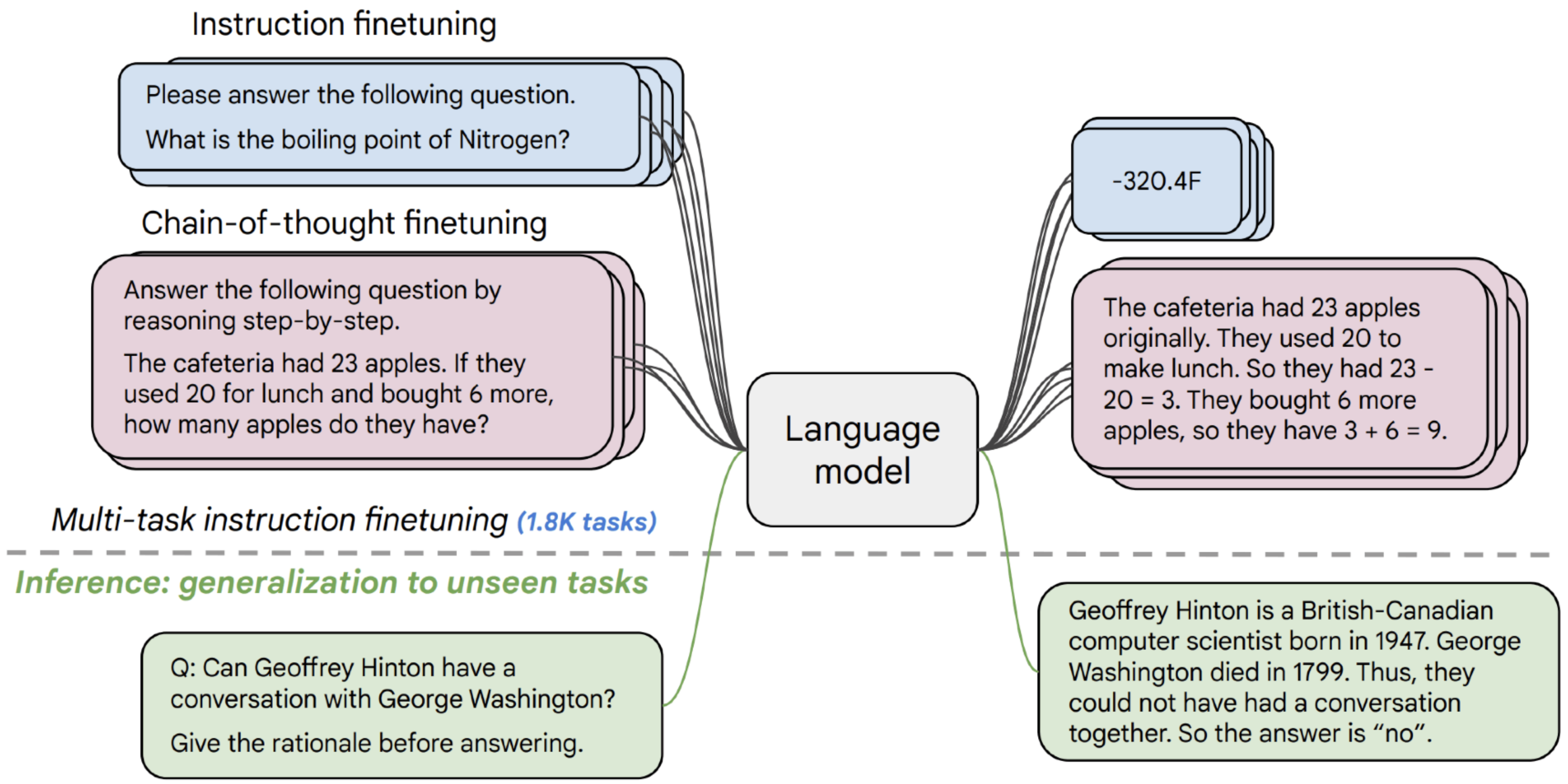
Generation as (Conditional) Language Modeling

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = p(y_1 \ldots y_n \mid x_1 \ldots x_m) = \prod_{t=1}^{n} p(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{x})$$



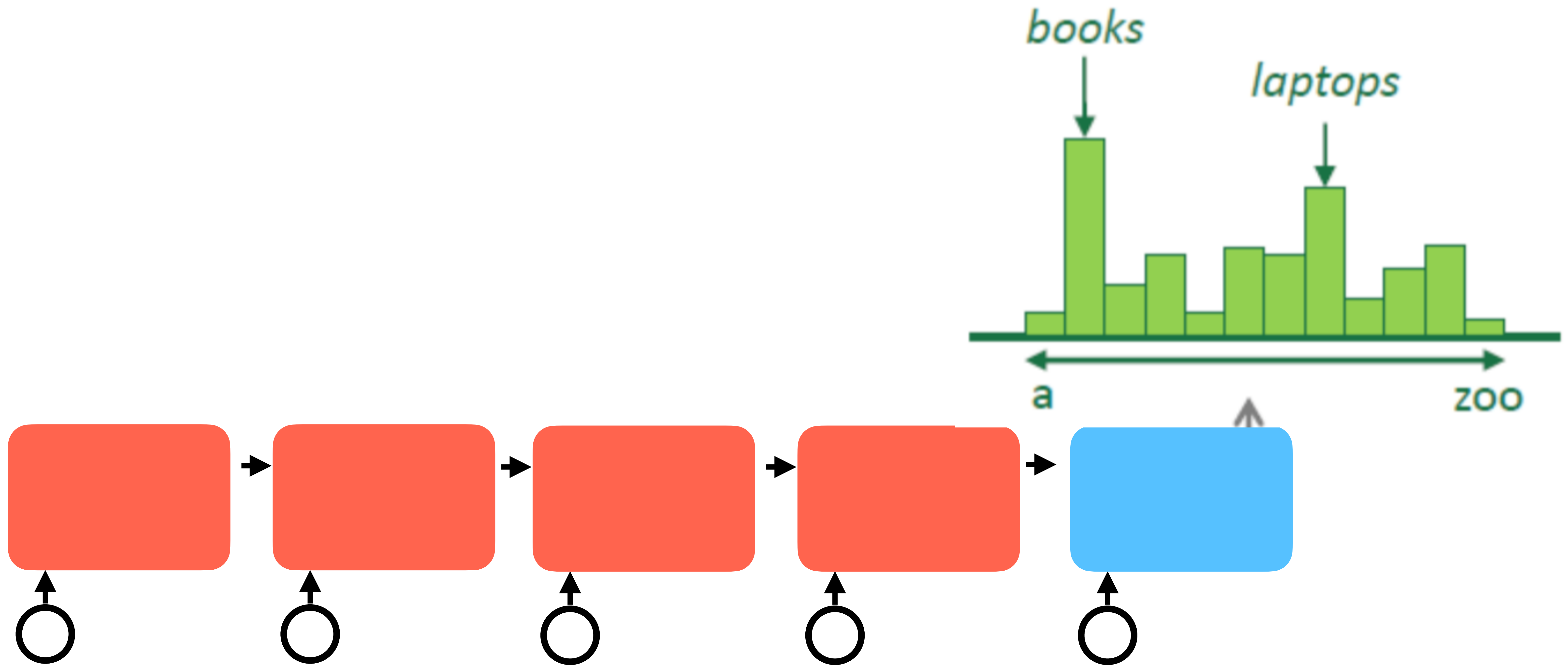
Continue Training / Post-Training / Instruction Tuning





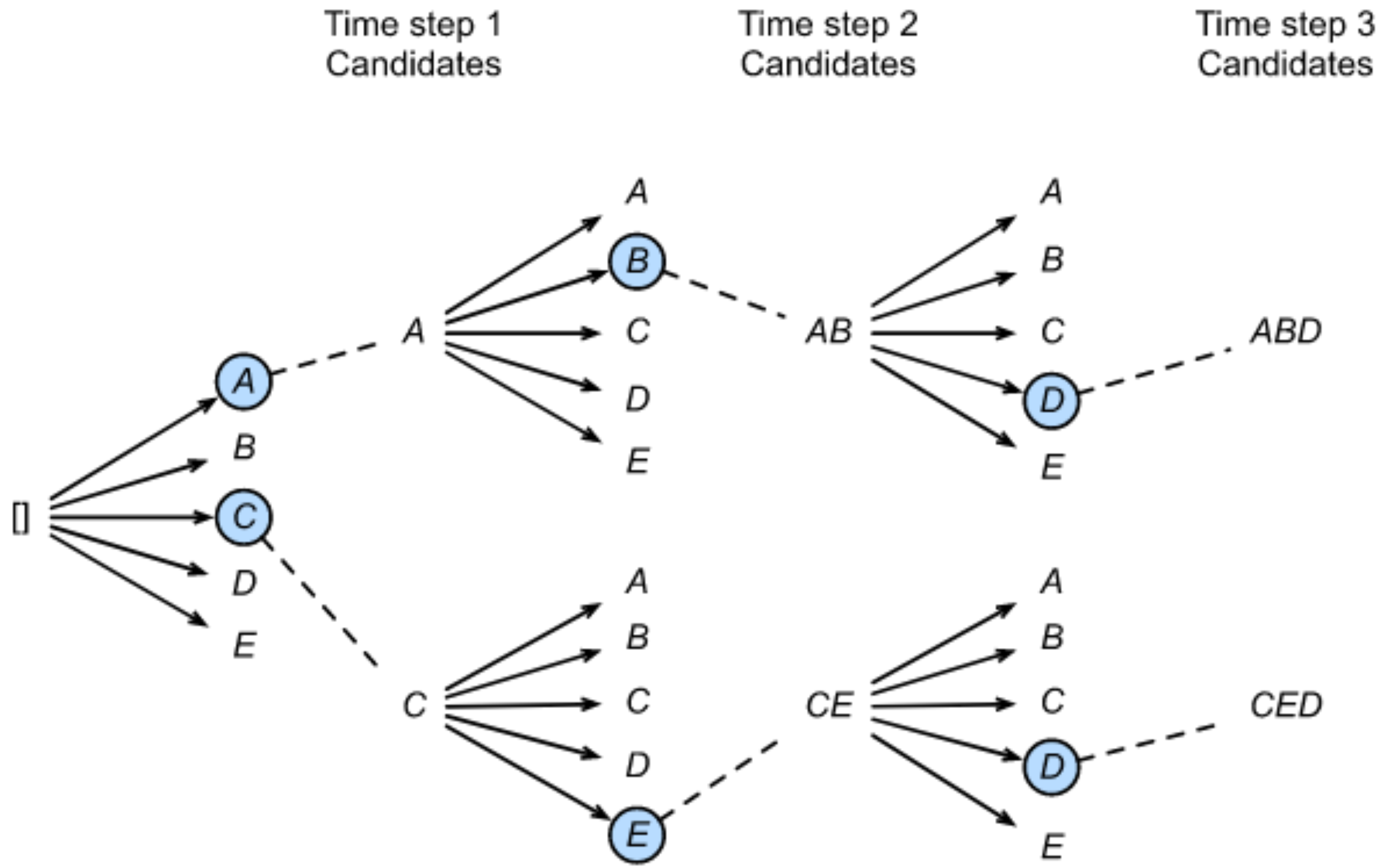
(Hyung Won Chung et al, 2022)

How to perform decoding?



$$p(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{x})$$

Beam Search



https://d2l.ai/
Dive into Depp Learning

Beam Search or Pure Sampling?

Context: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Beam Search, *b*=32:
"The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the Universidad Nacional Autónoma de México (UNAM) and the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de ..."

degenerate repetition

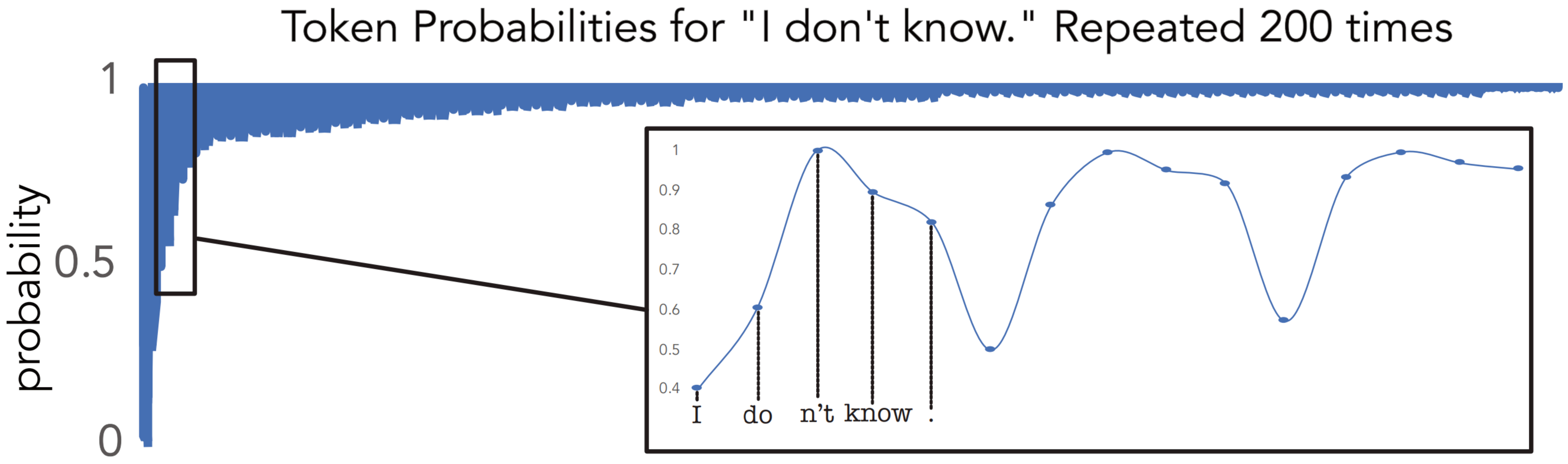
Pure Sampling:
They were cattle called Bolivian Cavalleros; they live in a remote desert uninterrupted by town, and they speak huge, beautiful, paradisiacal Bolivian linguistic thing. They say, 'Lunch, marge.' They don't tell what the lunch is," director Professor Chuperas Omwell told Sky News. "They've only been talking to scientists, like we're being interviewed by TV reporters. We don't even stick around to be interviewed by TV reporters. Maybe that's how they figured out that they're cosplaying as the Bolivian Cavalleros."

incoherent gibberish

GPT-2 Large (774M parameters)

Holtzman et al., 2020

What happened?



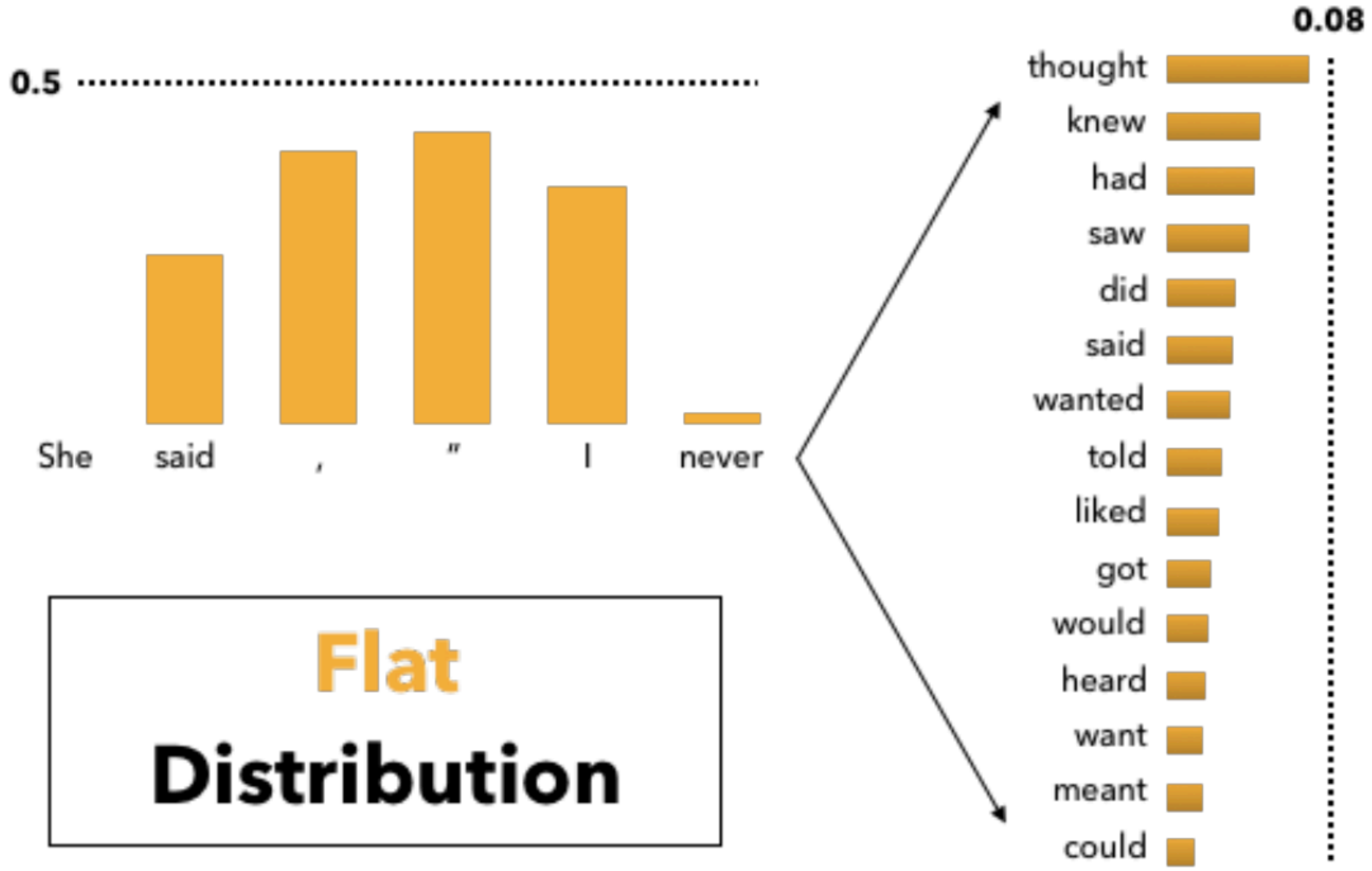
The probability of a repeated phrase increases with each repetition, creating a feedback loop.

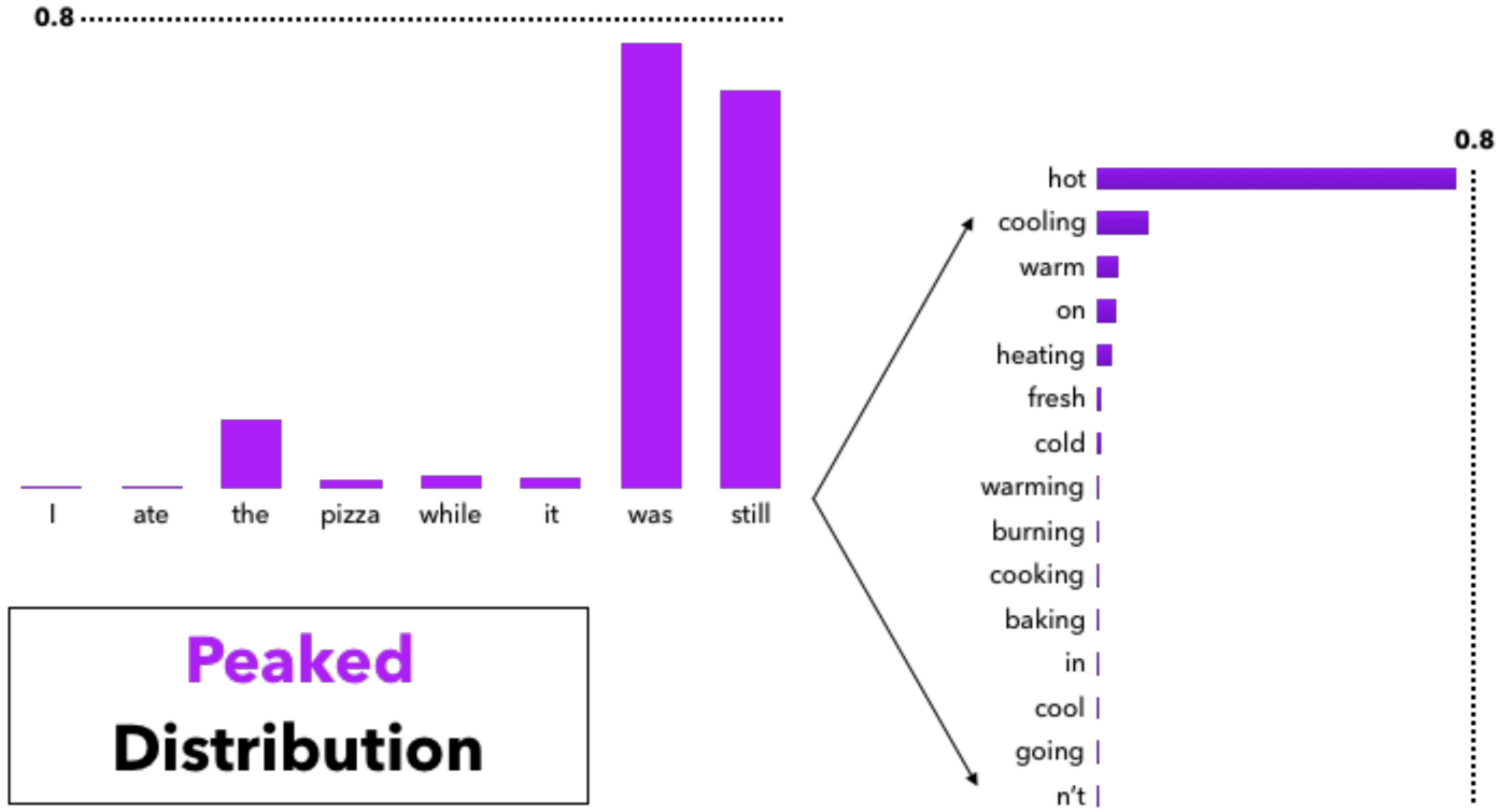
Holtzman et al., 2020

Top-K Sampling

Fan et al., 2018; Holtzman et al., 2020

Top-K Sampling

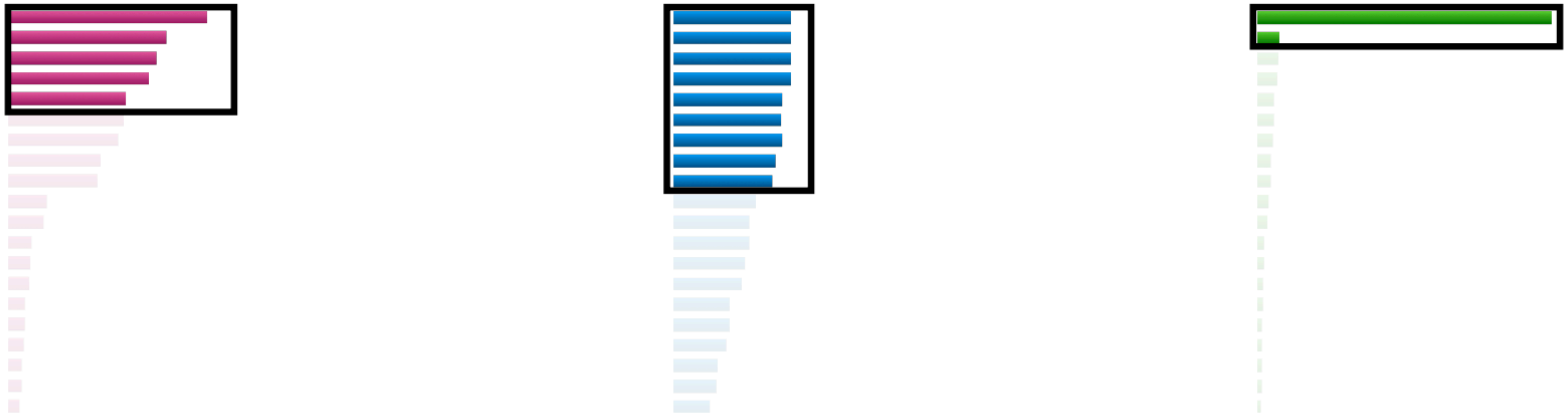




Holtzman et al., 2020

Top-p (nucleus) Sampling

To cut off by the cumulative probability mass, rather than the first K terms.



Holtzman et al., 2020