香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

COMP 4901B

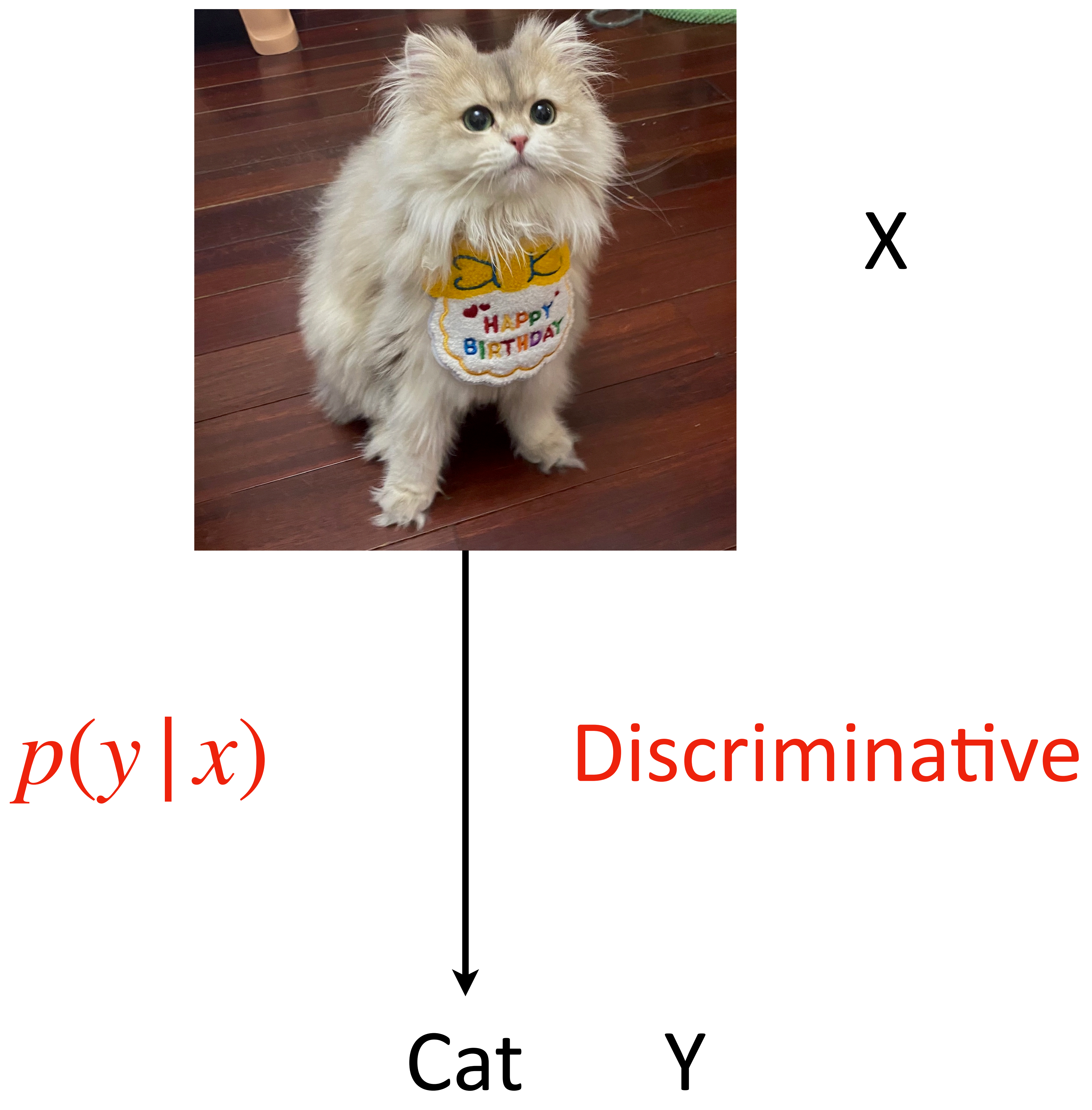
Large Language Models

Language Models

Junxian He

Sep 10, 2025

Discriminative vs. Generative Learning

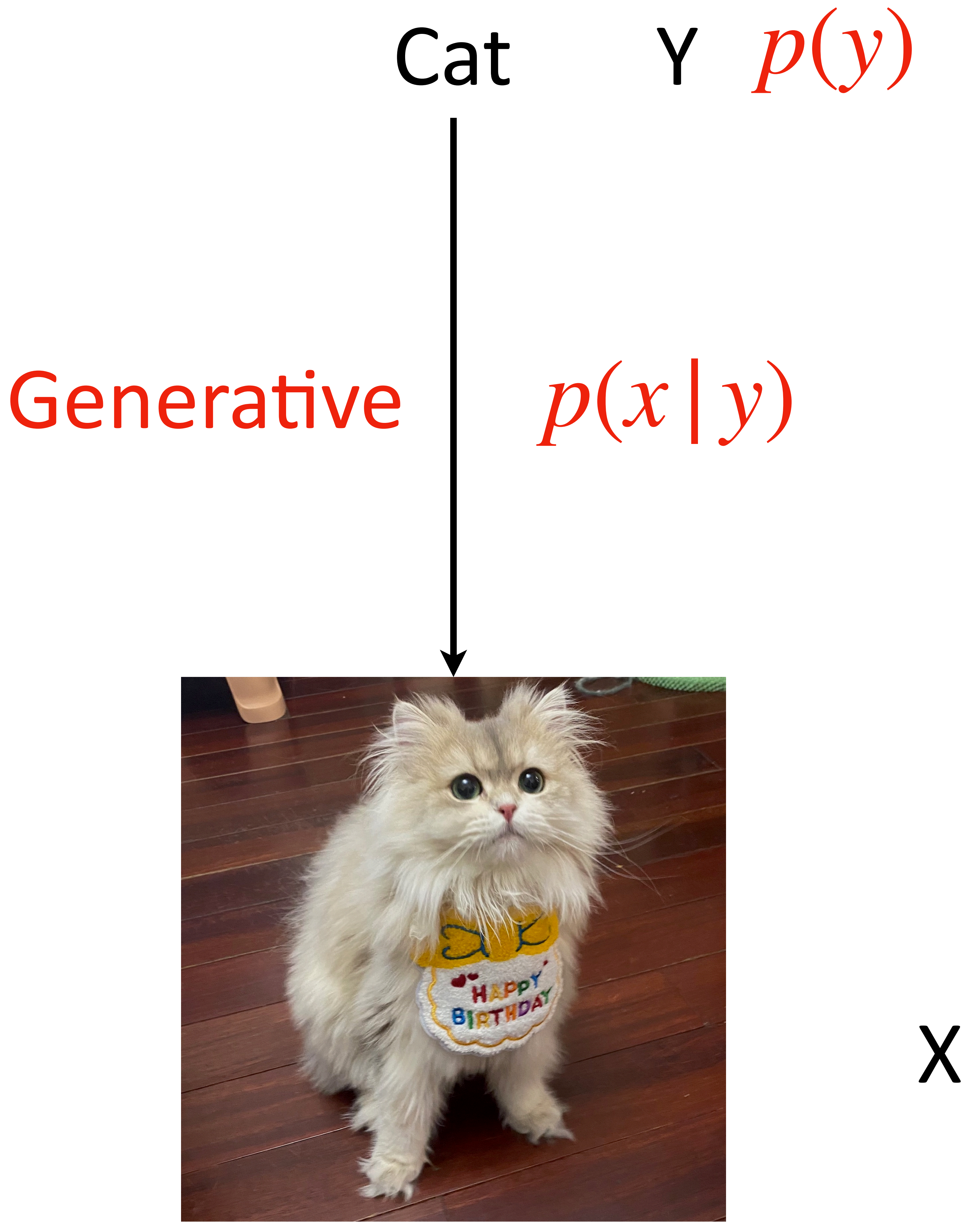


X

$p(y|x)$

Discriminative

Cat Y



Cat Y $p(y)$

Generative

$p(x|y)$

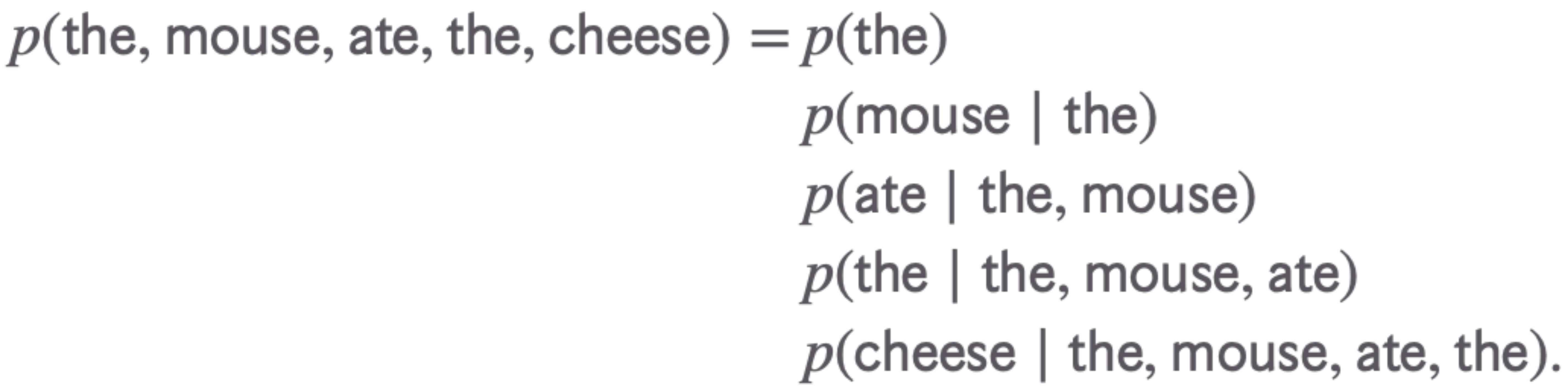
X

Probability of Sequences

Probability of multiple random variables:

$$p(x_1, x_2, \ldots, x_I) = \prod_{i=1}^{I} p(x_i | x_{1:i-1})$$

Probability of language:



$$\begin{aligned} p(\text{the, mouse, ate, the, cheese}) = p&(\text{the}) \\ p&(\text{mouse} \mid \text{the}) \\ p&(\text{ate} \mid \text{the, mouse}) \\ p&(\text{the} \mid \text{the, mouse, ate}) \\ p&(\text{cheese} \mid \text{the, mouse, ate, the}). \end{aligned}$$

Autoregressive language models

Autoregressive Language Models

$$\begin{aligned} p(\text{the, mouse, ate, the, cheese}) = {} & p(\text{the}) \\ & p(\text{mouse} \mid \text{the}) \\ & p(\text{ate} \mid \text{the, mouse}) \\ & p(\text{the} \mid \text{the, mouse, ate}) \\ & p(\text{cheese} \mid \text{the, mouse, ate, the}). \end{aligned}$$

$$p(x_1, x_2, \ldots, x_I) = \prod_{i=1}^{I} p(x_i \,|\, x_{1:i-1})$$

Next Word Context

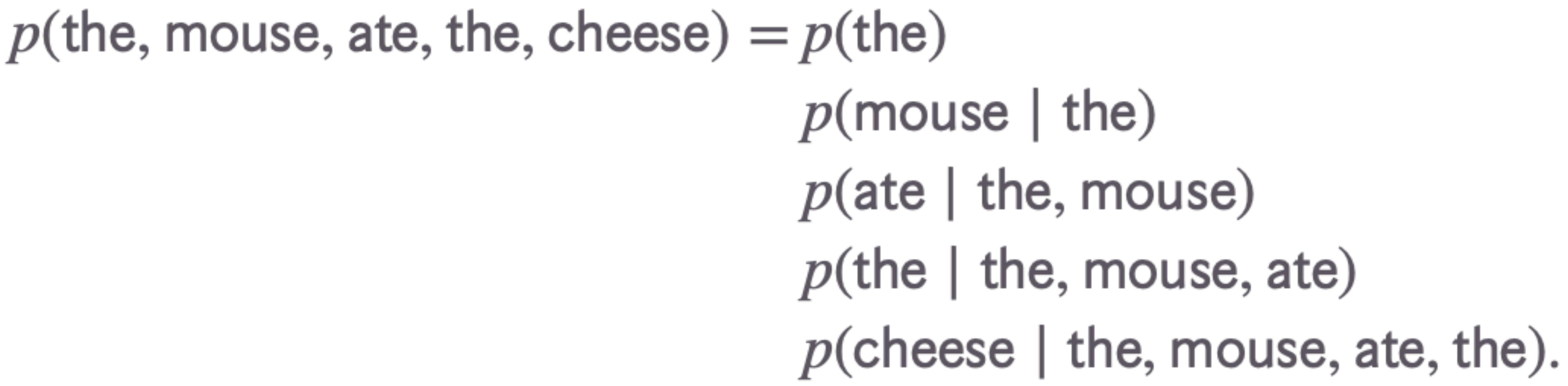
Autoregressive Language Models

$$\begin{aligned} p(\text{the, mouse, ate, the, cheese}) = {} & p(\text{the}) \\ & p(\text{mouse} \mid \text{the}) \\ & p(\text{ate} \mid \text{the, mouse}) \\ & p(\text{the} \mid \text{the, mouse, ate}) \\ & p(\text{cheese} \mid \text{the, mouse, ate, the}). \end{aligned}$$

$$p(x_1, x_2, \ldots, x_I) = \prod_{i=1}^{I} p(x_i \mid x_{1:i-1})$$

Learning a language model is to learn these conditional probabilities, for any language sequence

Autoregressive Language Models



$$\begin{aligned} p(\text{the, mouse, ate, the, cheese}) = {} & p(\text{the}) \\ & p(\text{mouse} \mid \text{the}) \\ & p(\text{ate} \mid \text{the, mouse}) \\ & p(\text{the} \mid \text{the, mouse, ate}) \\ & p(\text{cheese} \mid \text{the, mouse, ate, the}). \end{aligned}$$

$$p(x_1, x_2, \ldots, x_I) = \prod_{i=1}^{I} p(x_i \,|\, x_{1:i-1})$$

Given a dataset, how to find these probabilities?

Maximum Likelihood Estimation

Count-based Language Models

Count the frequency and divide

$$p(x_i \mid x_{1:i-1}) = \frac{c(x_{1:i})}{c(x_{1:i-1})}$$

There are infinite number of parameters for language

We may see long sequences only once, counting becomes meaningless

n-gram Language Models

Next token probability only depends on the previous n-1 words

Unigram LM:

$$p(x_1, x_2, \ldots, x_I) = \prod_{i=1}^{I} p(x_i)$$

Each token is independent

Bigram LM:

$$p(x_1, x_2, \ldots, x_I) = \prod_{i=1}^{I} p(x_i \,|\, x_{i-1})$$

Generally for n-gram LM:

$$p(x_1, x_2, \ldots, x_I) = \prod_{i=1}^{I} p(x_i \,|\, x_{i-n+1:i-1})$$

Parameter Estimation for n-gram LM

Count-based:

$$p(x_i \mid x_{i-n+1:i-1}) = \frac{c(x_{i-n+1:i})}{c(x_{i-n+1:i-1})}$$

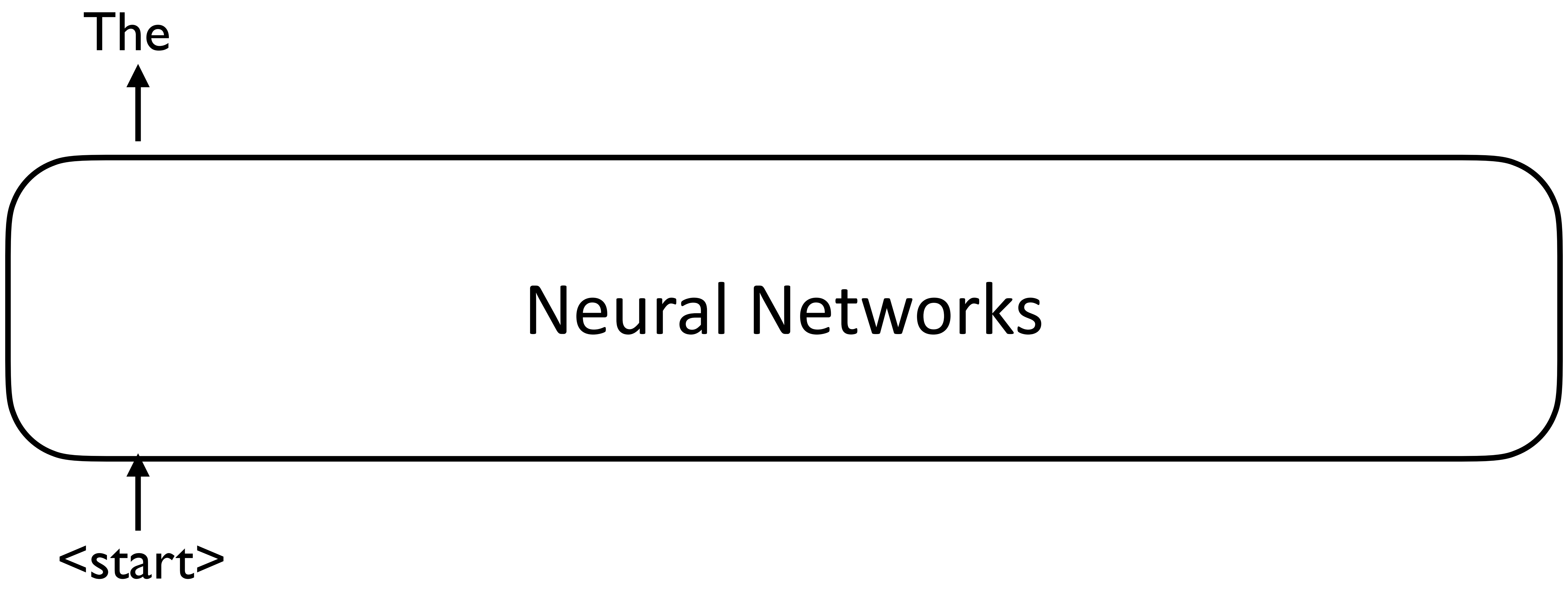
Number of parameters decreases, but flexibility decreases as well

Traditionally, we directly compute this probability, but neural language models use neural networks to compute the probability

Neural Language Models

Neural language models are typically autoregressive

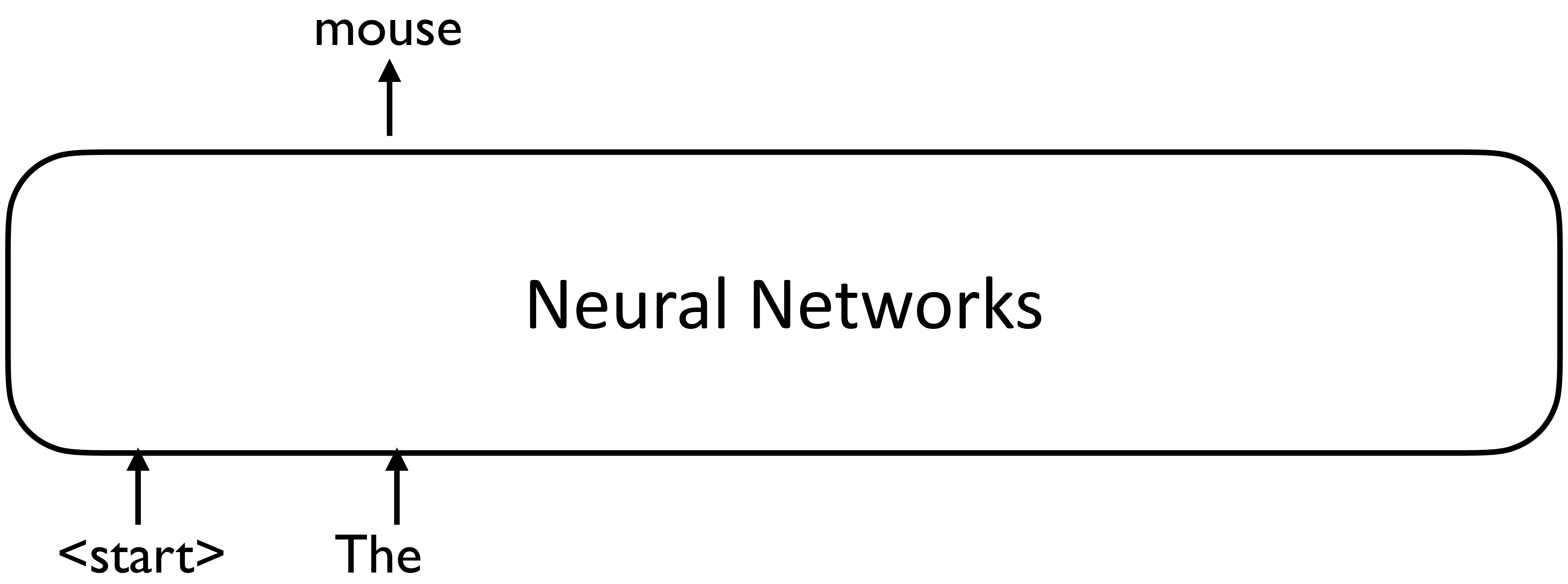
Data: "The mouse ate the cheese ."



Neural Language Models

Neural language models are typically autoregressive

Data: "The mouse ate the cheese ."



mouse

Neural Networks

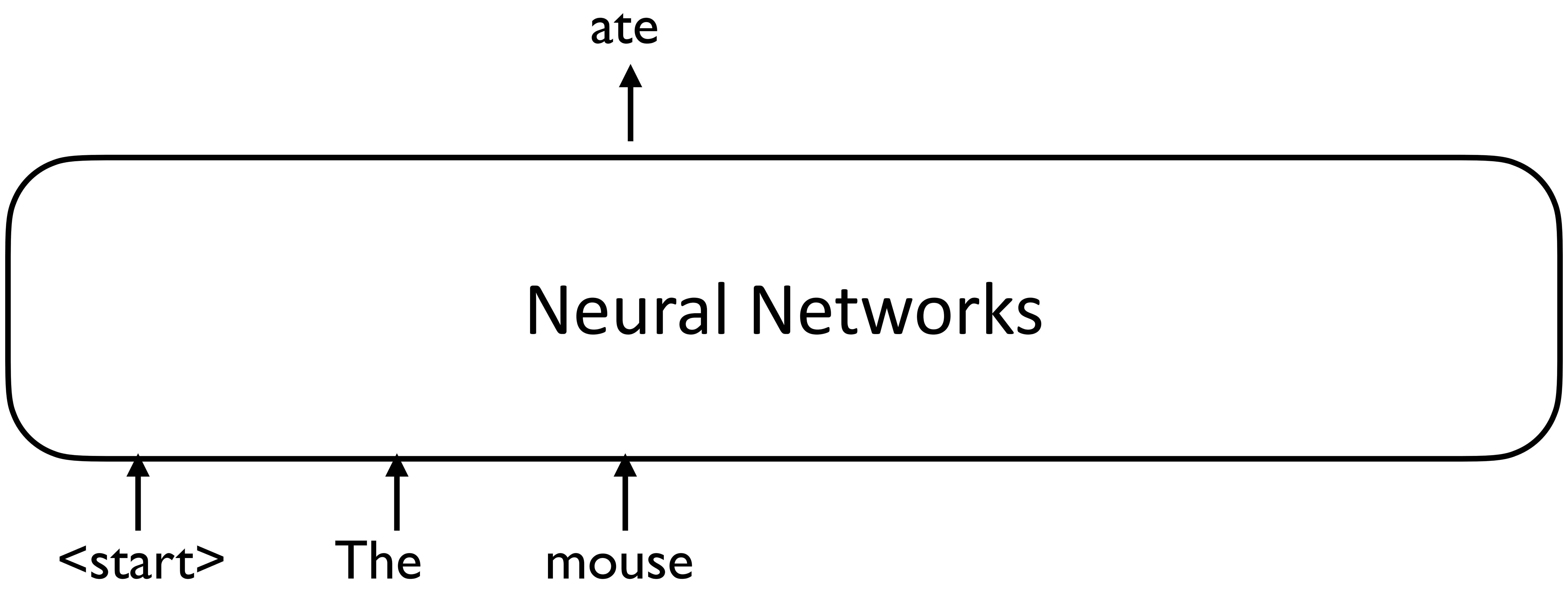
<start>

The

Neural Language Models

Neural language models are typically autoregressive

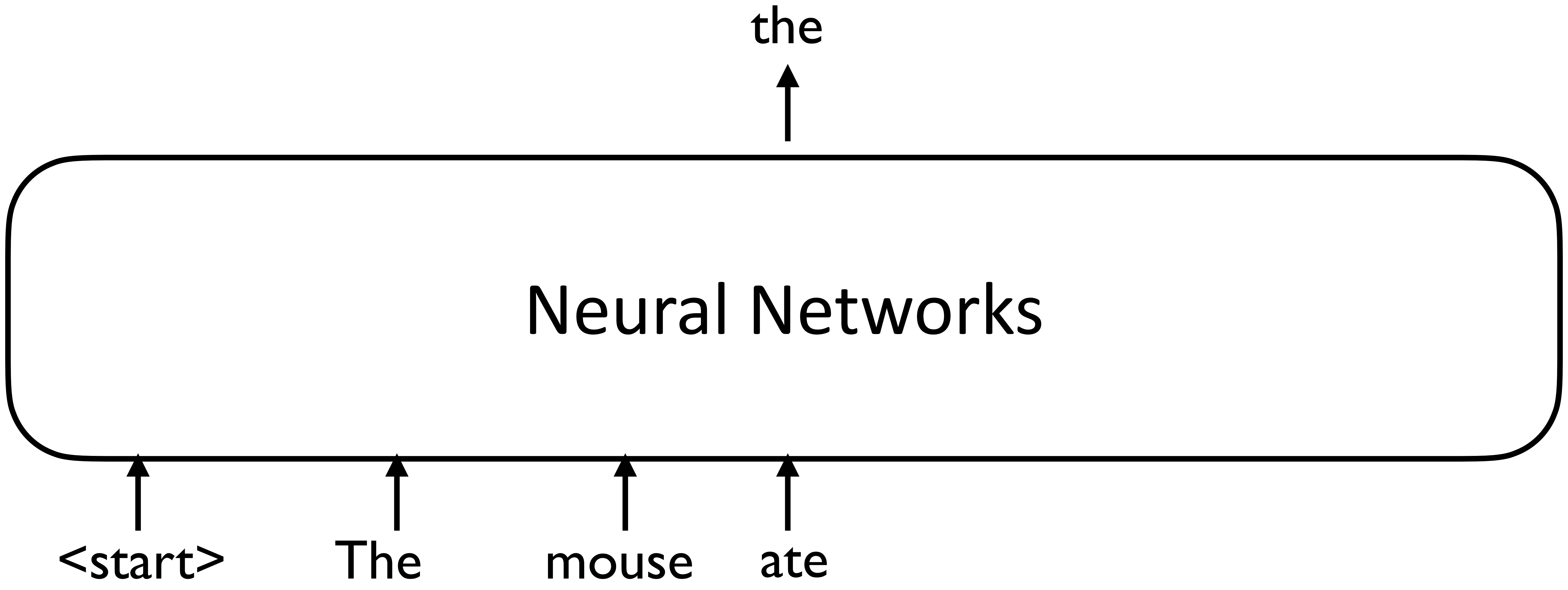
Data: “The mouse ate the cheese .”



Neural Language Models

Neural language models are typically autoregressive

Data: "The mouse ate the cheese ."

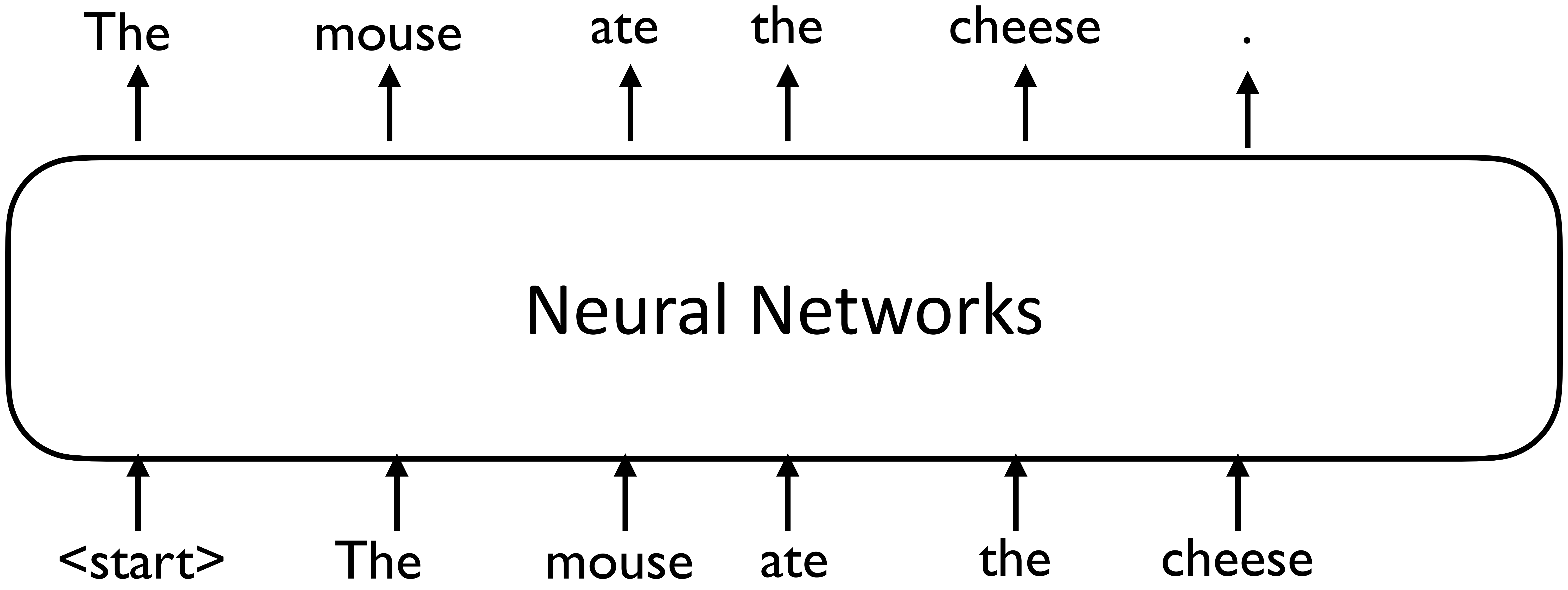


We can compute the loss on every token in parallel

Neural Language Models

Neural language models are typically autoregressive

Data: "The mouse ate the cheese ."



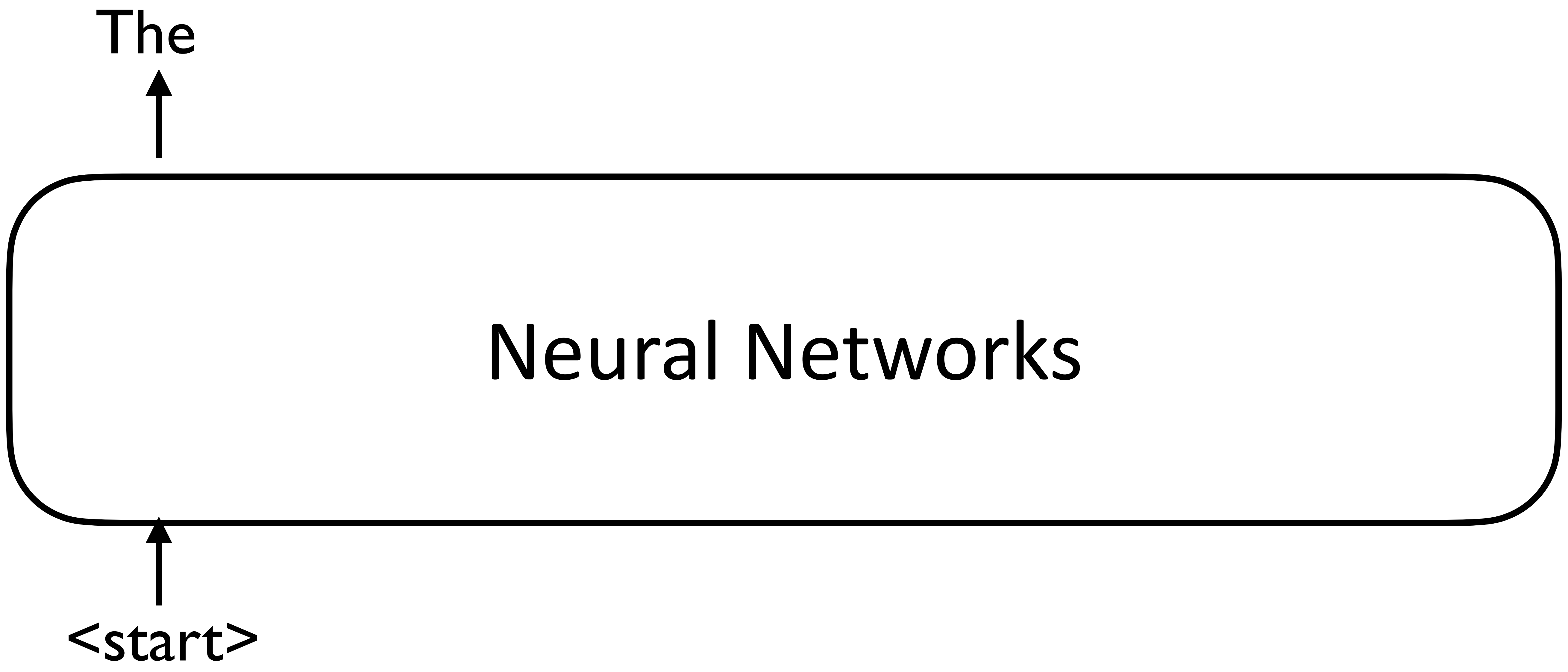
Each prediction only sees the inputs on its left

Neural Language Models

Are language models generative models? ✅

Can we compute p(x) given x? Can we sample new x? ✅

At inference time, to generate:

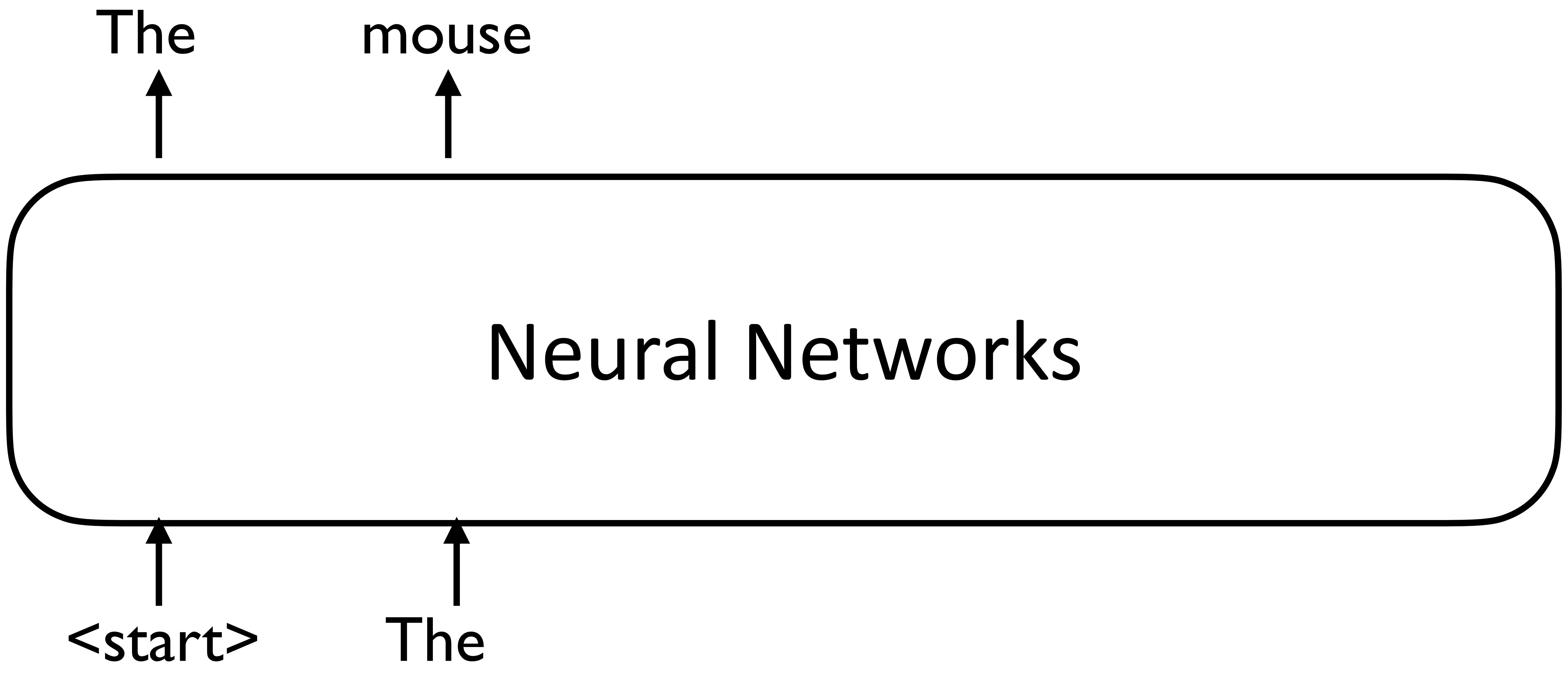


Neural Language Models

Are language models generative models? ✅

Can we compute p(x) given x? Can we sample new x? ✅

At inference time, to generate:

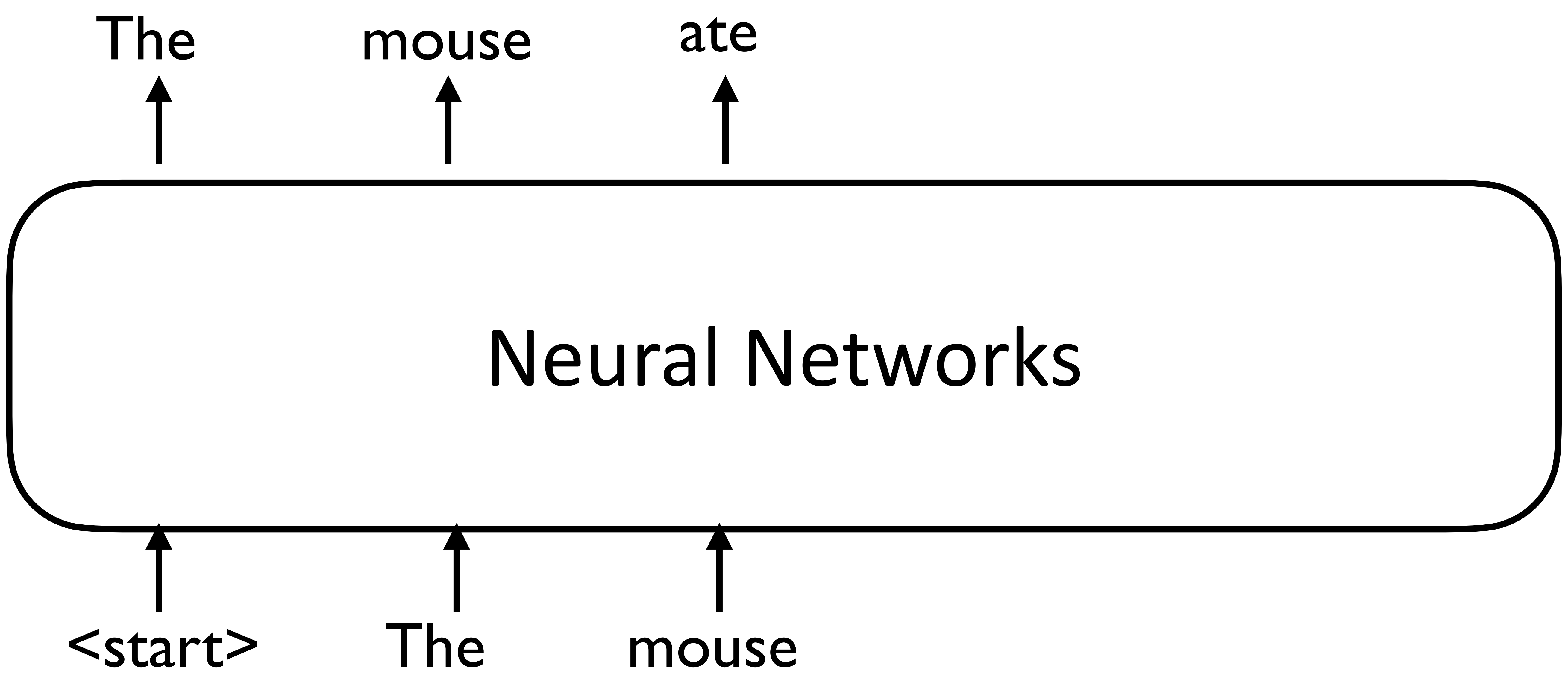


Neural Language Models

Are language models generative models? ✅

Can we compute p(x) given x? Can we sample new x? ✅

At inference time, to generate:

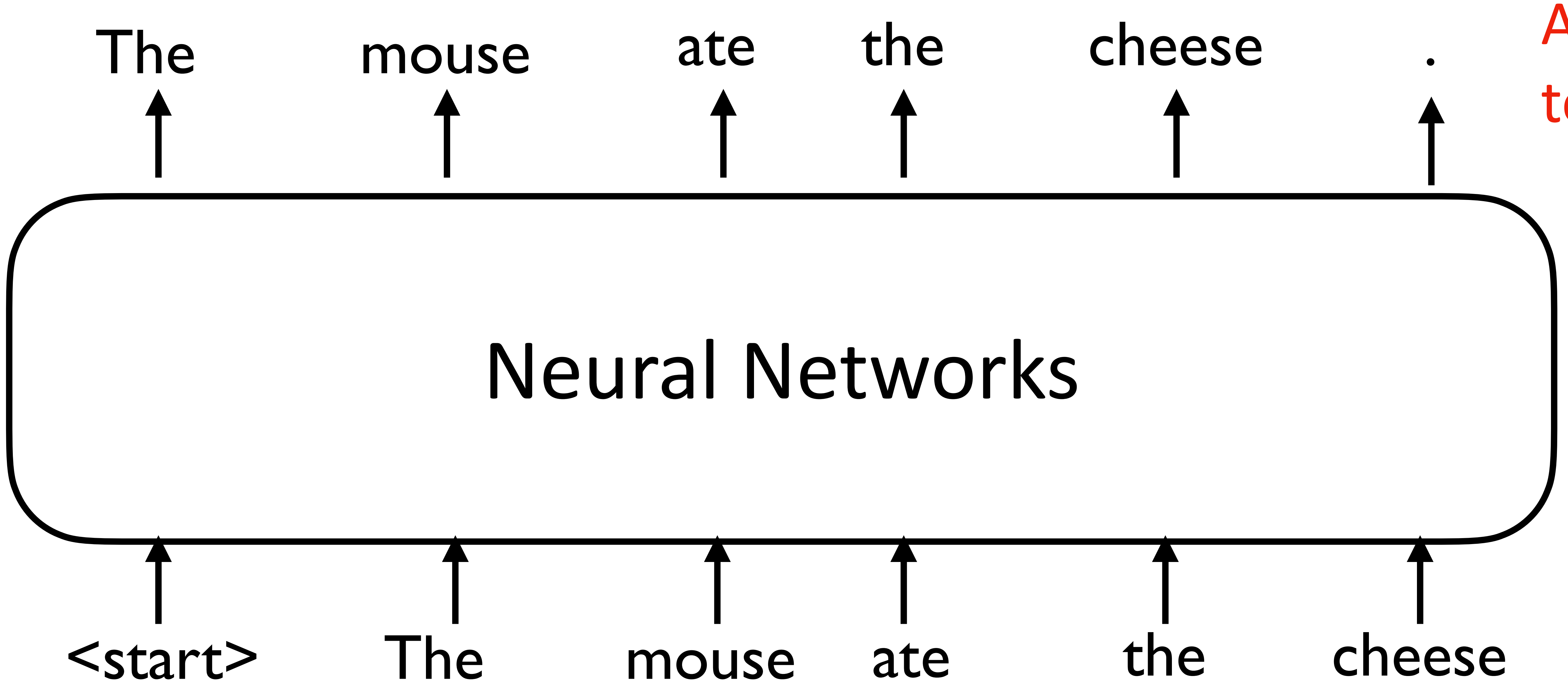


Neural Language Models

Are language models generative models? ✅

Can we compute p(x) given x? Can we sample new x? ✅

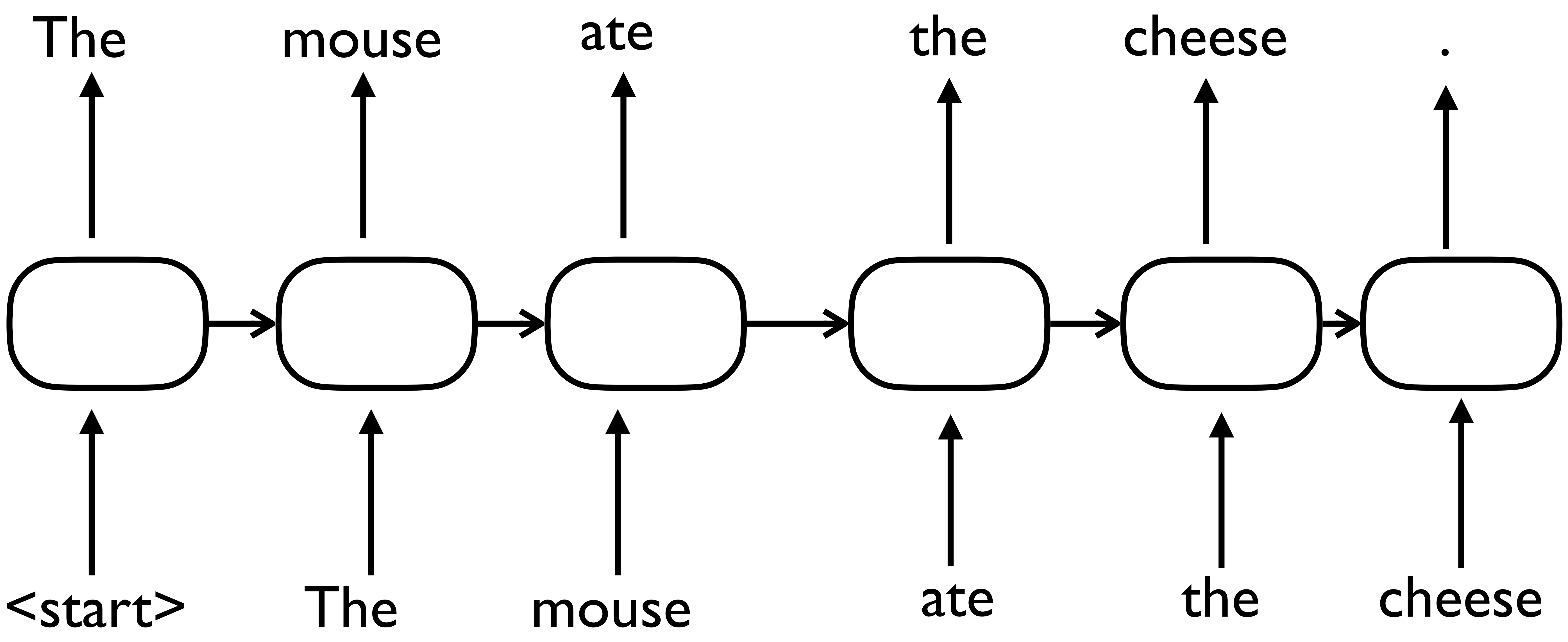
At inference time, to generate:



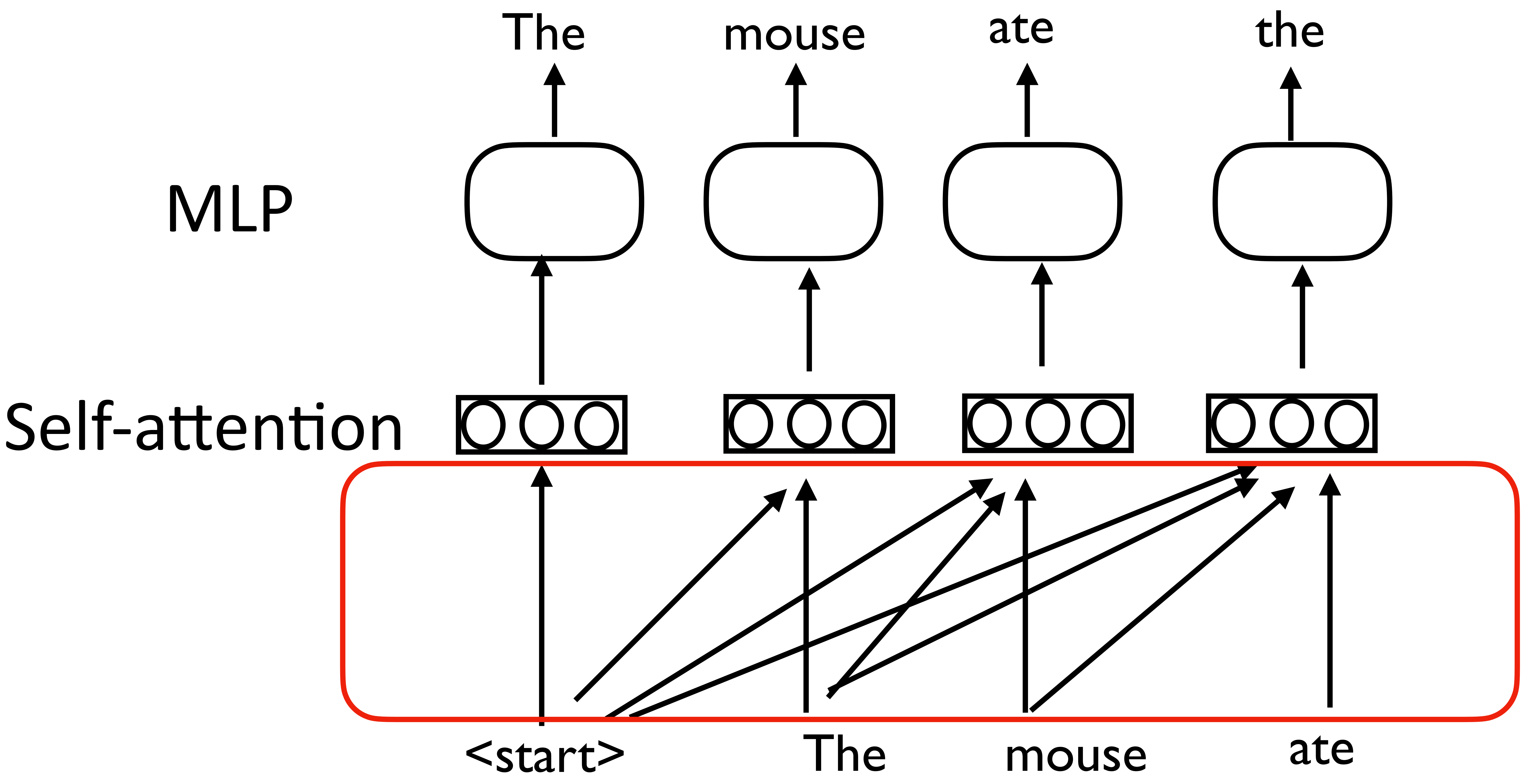
Autoregressive generation has to generate token by token

Cann't parallelize, efficiency of autoregressive decoding is still an important research topic

RNN Language Models



Transformer Language Models

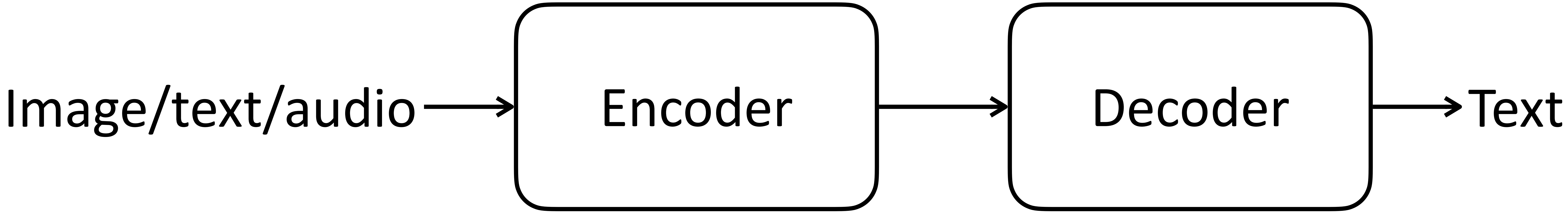


Self-attention only attends to the tokens on the left (masked attention)

Neural Language Models

Language model is the fundamental block to model language distribution p(x)

For a long time, to solve specific tasks:



When we have a better arch/training for LM, we can have a better decoder

Not long ago, some people think purely language models is useless because it does not directly address tasks, and LM performance may not transfer to downstream tasks

Is Next Token Prediction Useful?

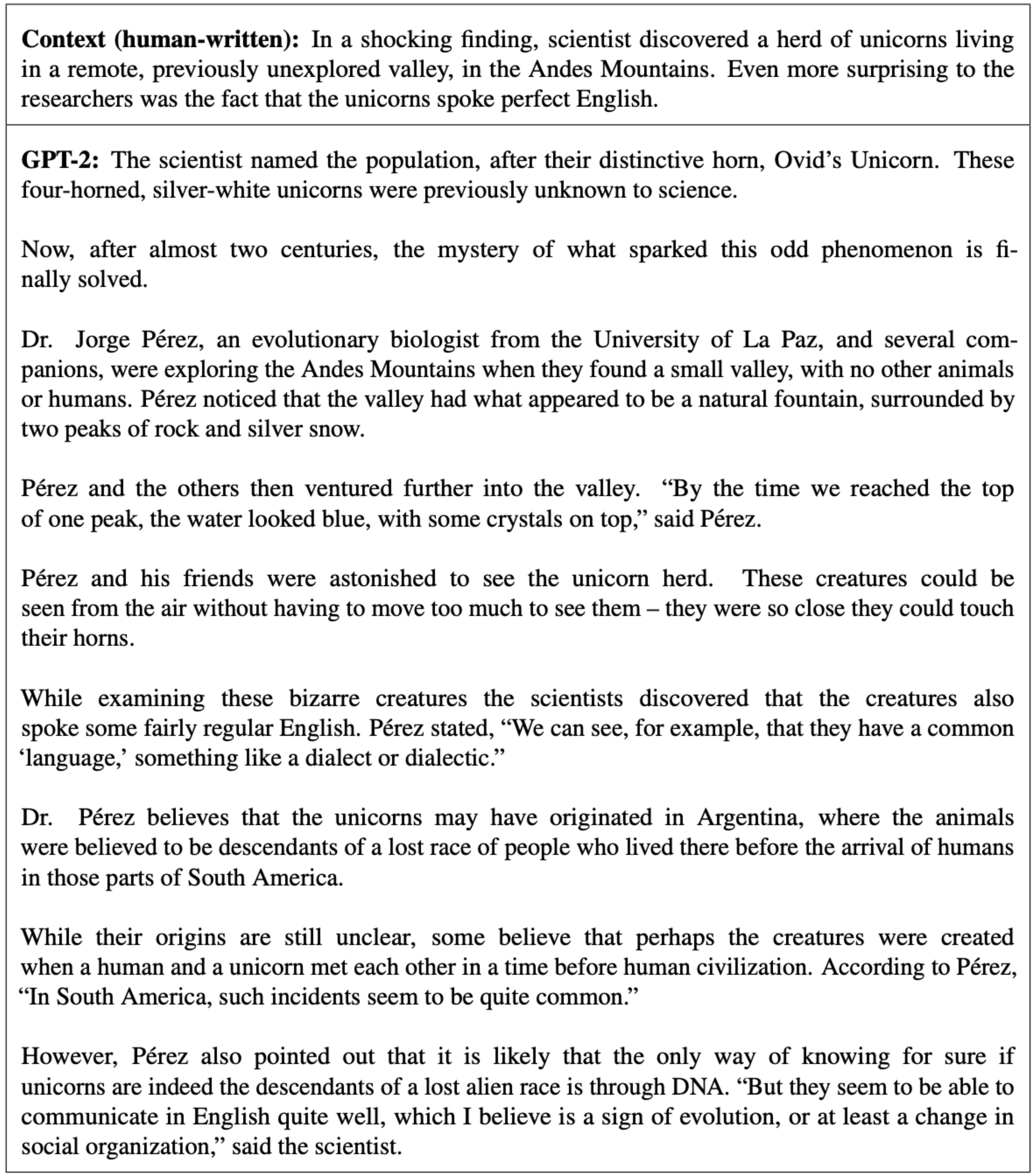
Ok, language modeling can be used as pretraining, but is a language model itself useful for some tasks directly?

In the late 1980s the Hong Kong Government anticipated a strong demand for university graduates to fuel an economy increasingly based on services. Sir Sze-Yuen Chung and Sir Edward Youde, the then Governor of Hong Kong, conceived the idea of another university in addition to the pre-existing two universities, The University of Hong Kong and The Chinese University of Hong Kong.
Planning for the "Third University", named The Hong Kong University of Science and Technology later, began in 1986. Construction began at the Kohima Camp site in Tai Po Tsai on the Clear Water Bay Peninsula. The site was earmarked for the construction of a new [ ]

Completion

This task seems useless in practice

Language Models are Zero-Shot Learners



Context (human-written): 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.

GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, "In South America, such incidents seem to be quite common."

However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. "But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization," said the scientist.

Radford et al. Language Models are Unsupervised Multitask Learners. 2019 (GPT2)

GPT-2

Next token prediction can unify many tasks

Machine translation:

Chinese: 今天是学期的最后一天。
English:

Completion is very general

This was an early form of prompting, that is widely discussed today

Question answering:

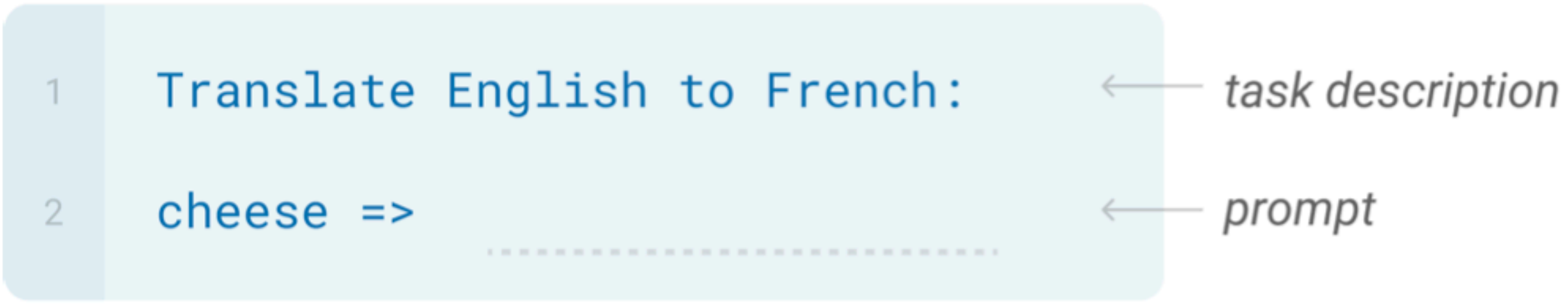
Q: What is the capital of the United States?
A:

Radford et al. Language Models are Unsupervised Multitask Learners. 2018.

Language Models Are Few-Shot Learners

Zero-shot

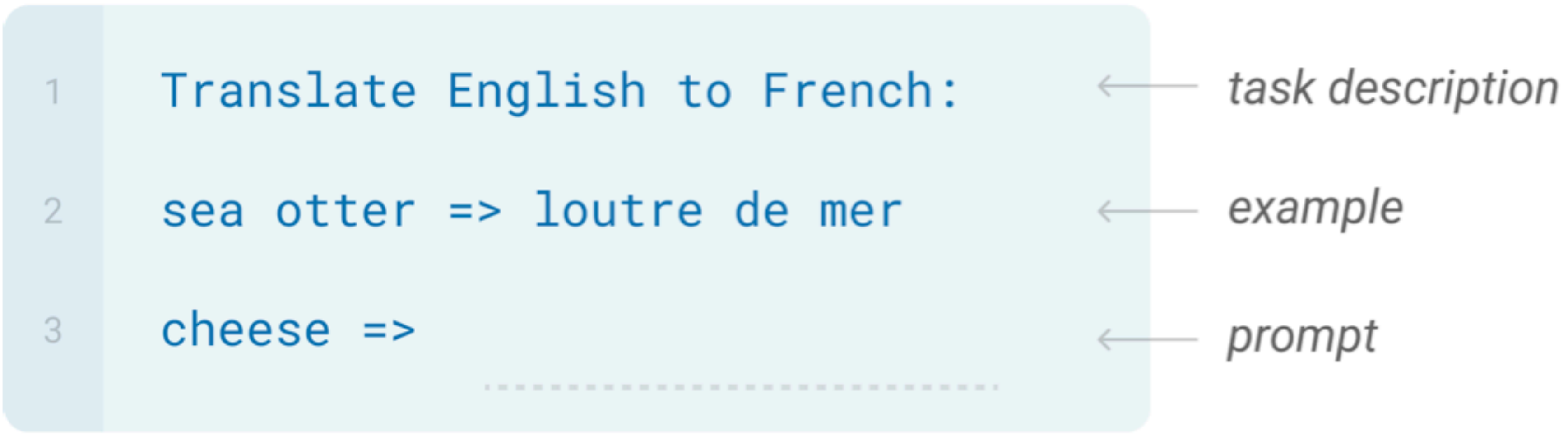
The model predicts the answer given only a natural language description of the task. No gradient updates are performed.



```
Translate English to French:
cheese =>
```

One-shot

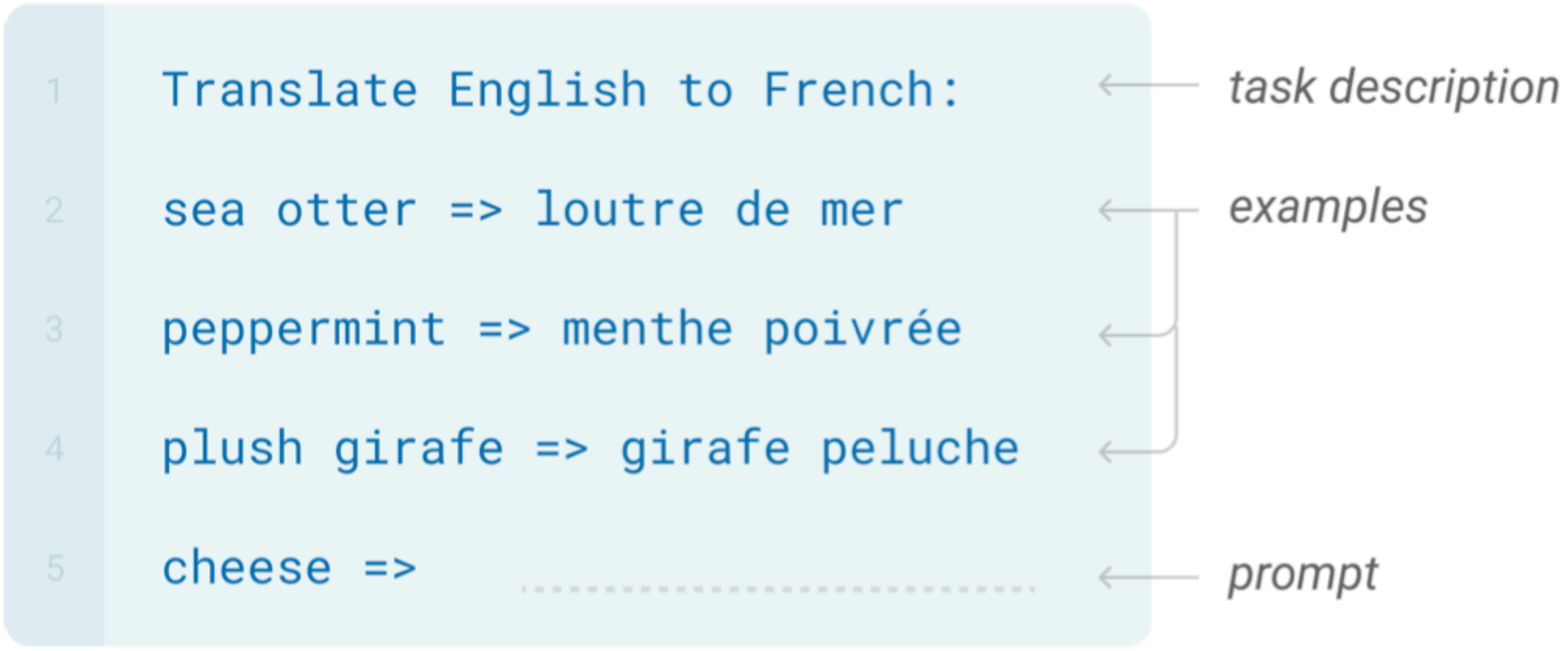
In addition to the task description, the model sees a single example of the task. No gradient updates are performed.



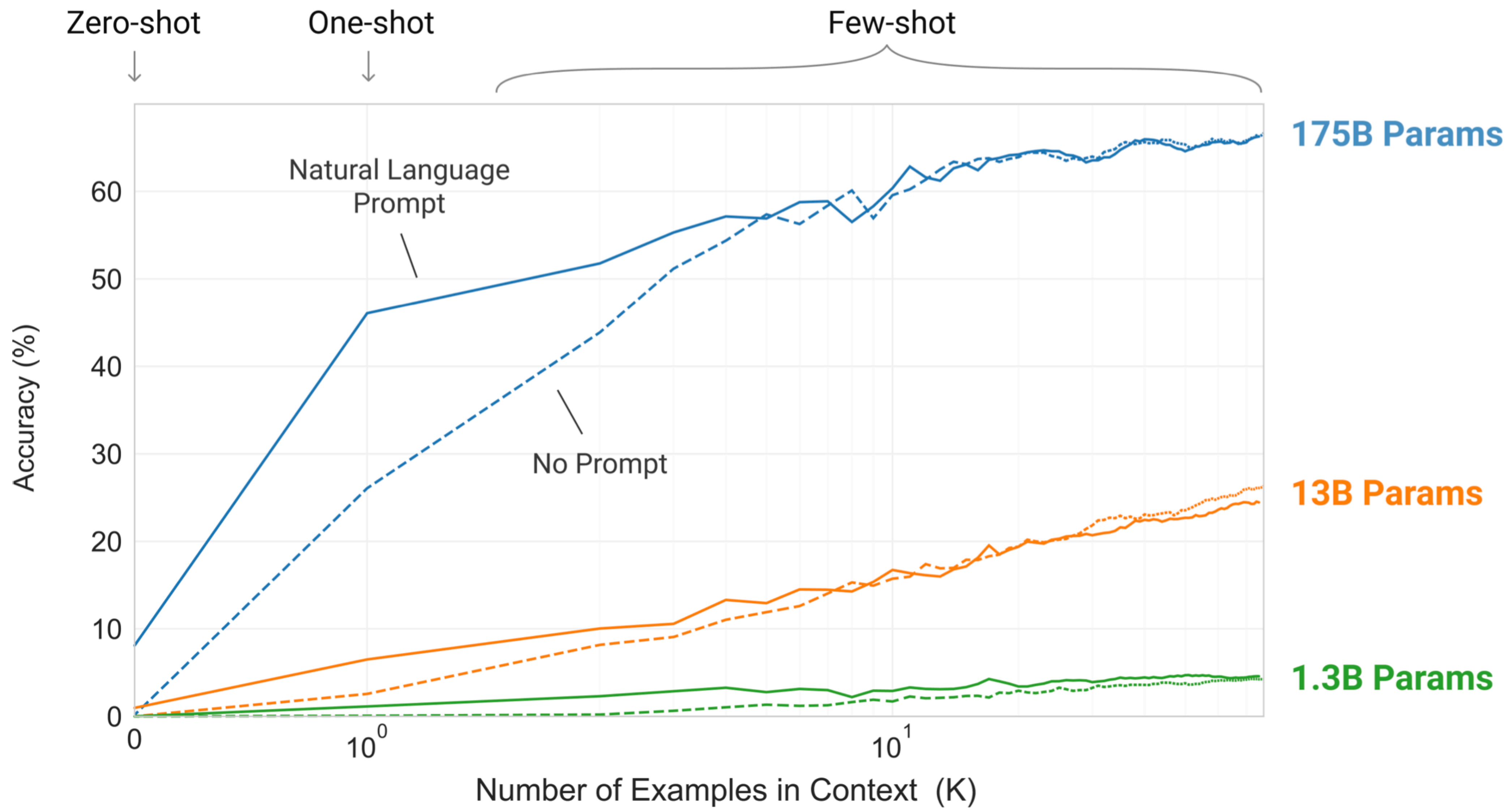
```
Translate English to French:
sea otter => loutre de mer
cheese =>
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



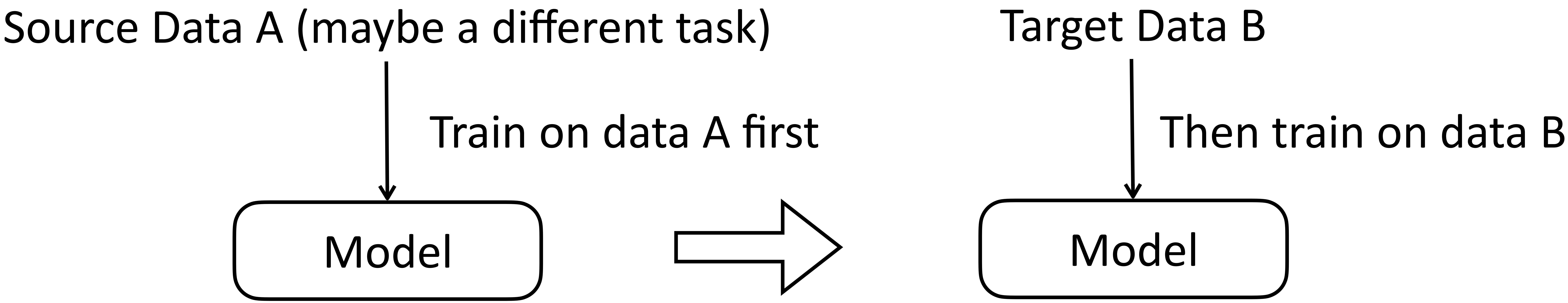
```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```



In-Context Learning

Brown et al. Language models are few-shot learners. 2020

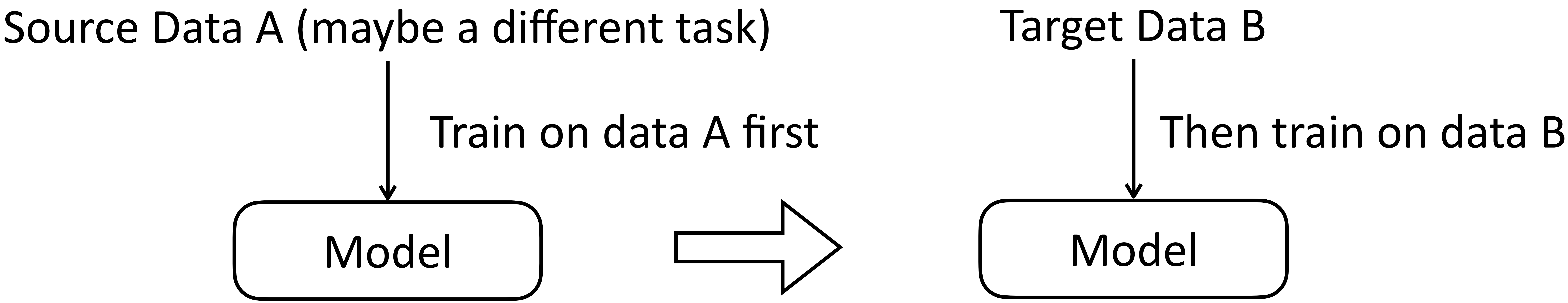
Pretraining



Classically, this is transfer Learning

It is now called pretraining because of the scale of A

Pretraining



Source Data A (maybe a different task)

Train on data A first

Model

Target Data B

Then train on data B

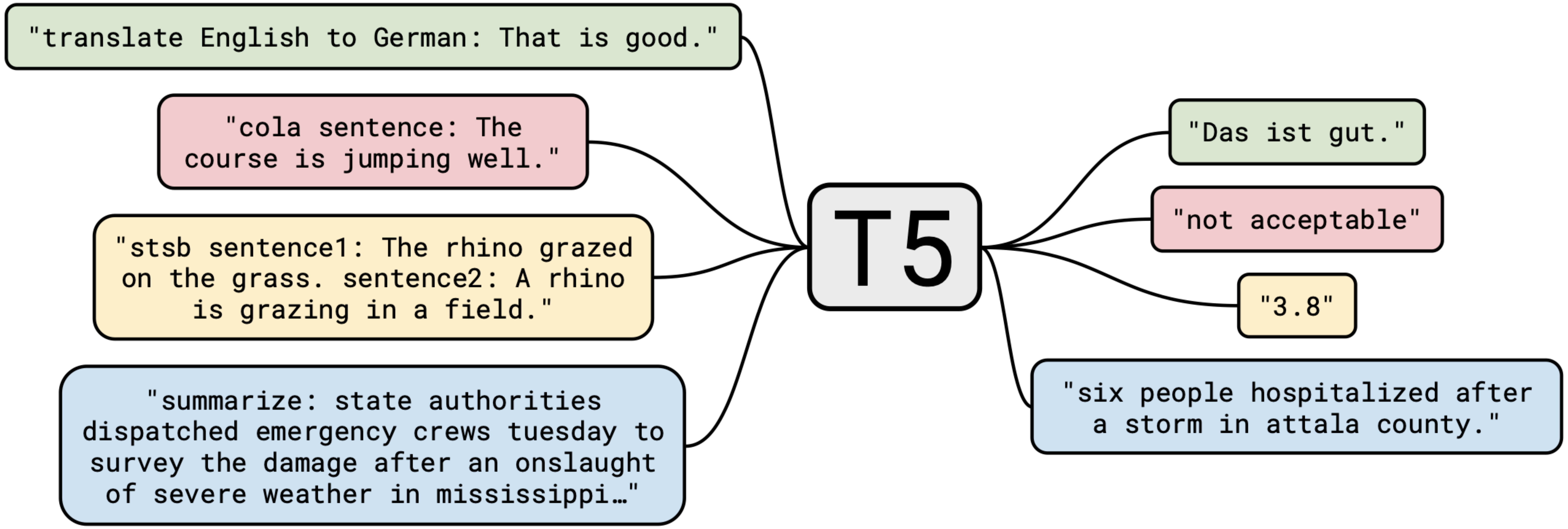
Model

For supervised training, data A is often limited

How can we find large-scale data A to train?

Prompt Breaks Task Boundaries

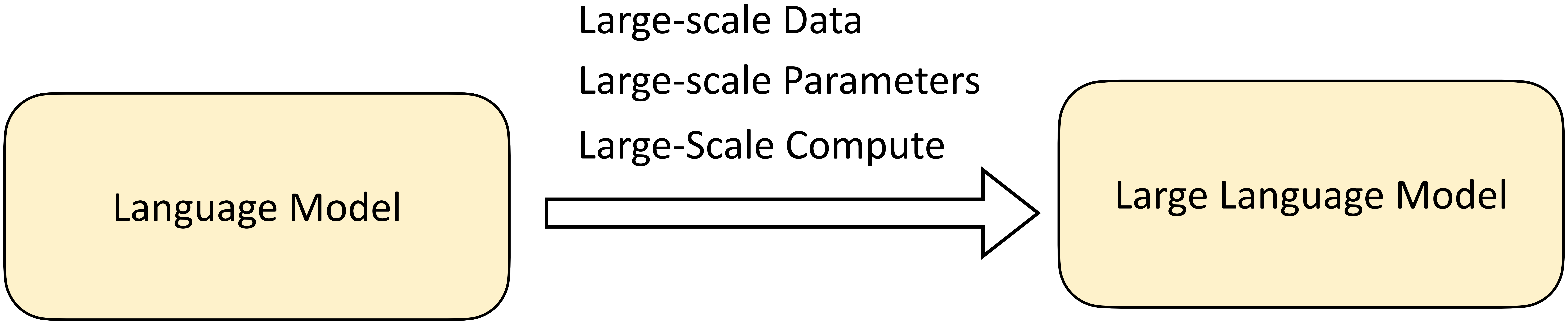
Almost all text tasks can be expressed with a unified format, no matter whether it is classification or generation



"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

Raffle et al. Exploring the Limits of Transfer Learning. 2020

Large Language Models



Language Model

Large-scale Data

Large-scale Parameters

Large-Scale Compute

Large Language Model

Thank You!