# EE542
## Lecture 23: Large Language Model Part 2

Internet and Cloud Computer

Young Cho
Department of Electrical Engineering
University of Southern California

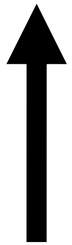


*Based on LLM Slides from Steve Seitz with permission*

Early one morning the sun was shining I was laying in bed

Wondering if she had changed at all if her hair was still red

red



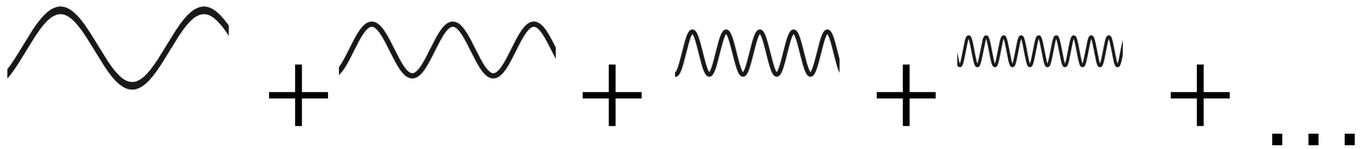
Early one morning the sun was shining I was laying in bed

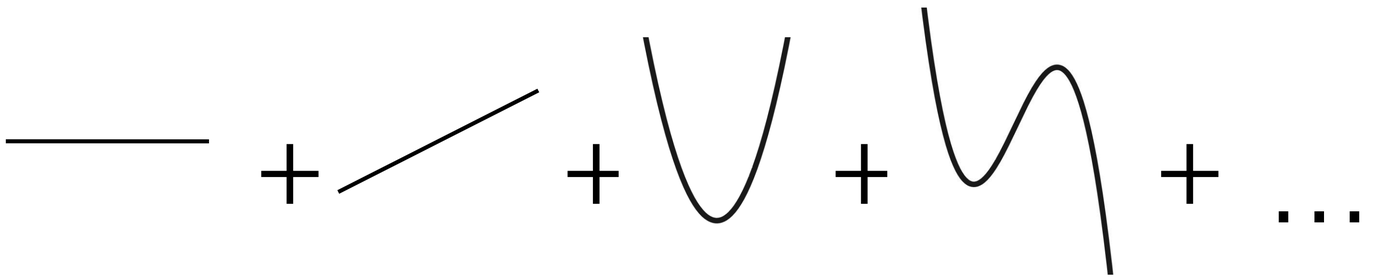
Wondering if she had changed at all if her hair was still

$$P(x_n|x_{n-1}, x_{n-2}, x_{n-3}, x_{n-4}, x_{n-5}, x_{n-6}, x_{n-7}, x_{n-8}, x_{n-9}, x_{n-10}, x_{n-11}, x_{n-12}, x_{n-13})$$

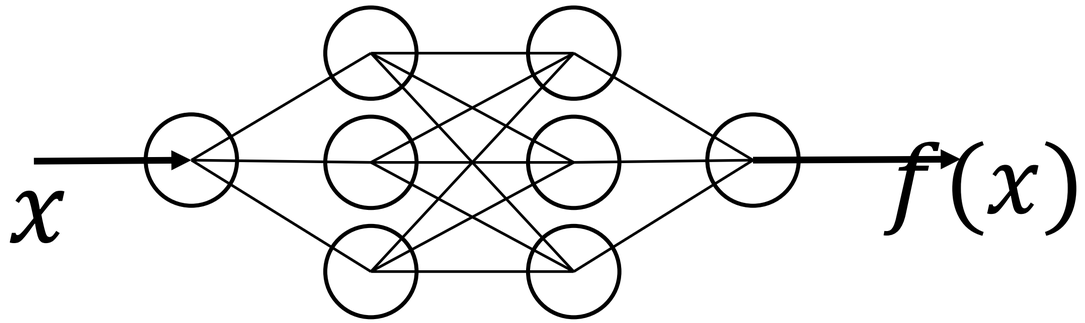
$10^{70}$combinations

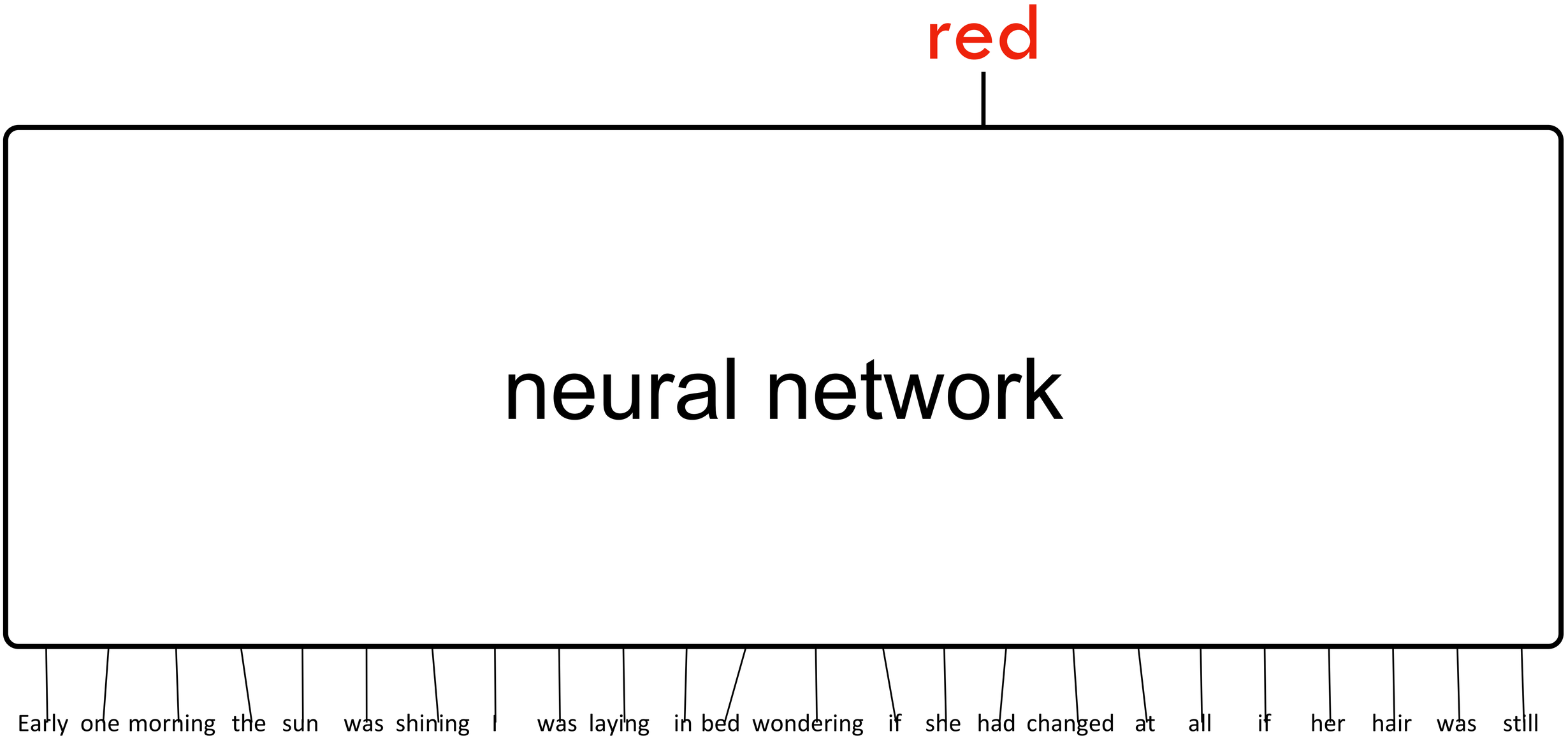
# Function Approximation

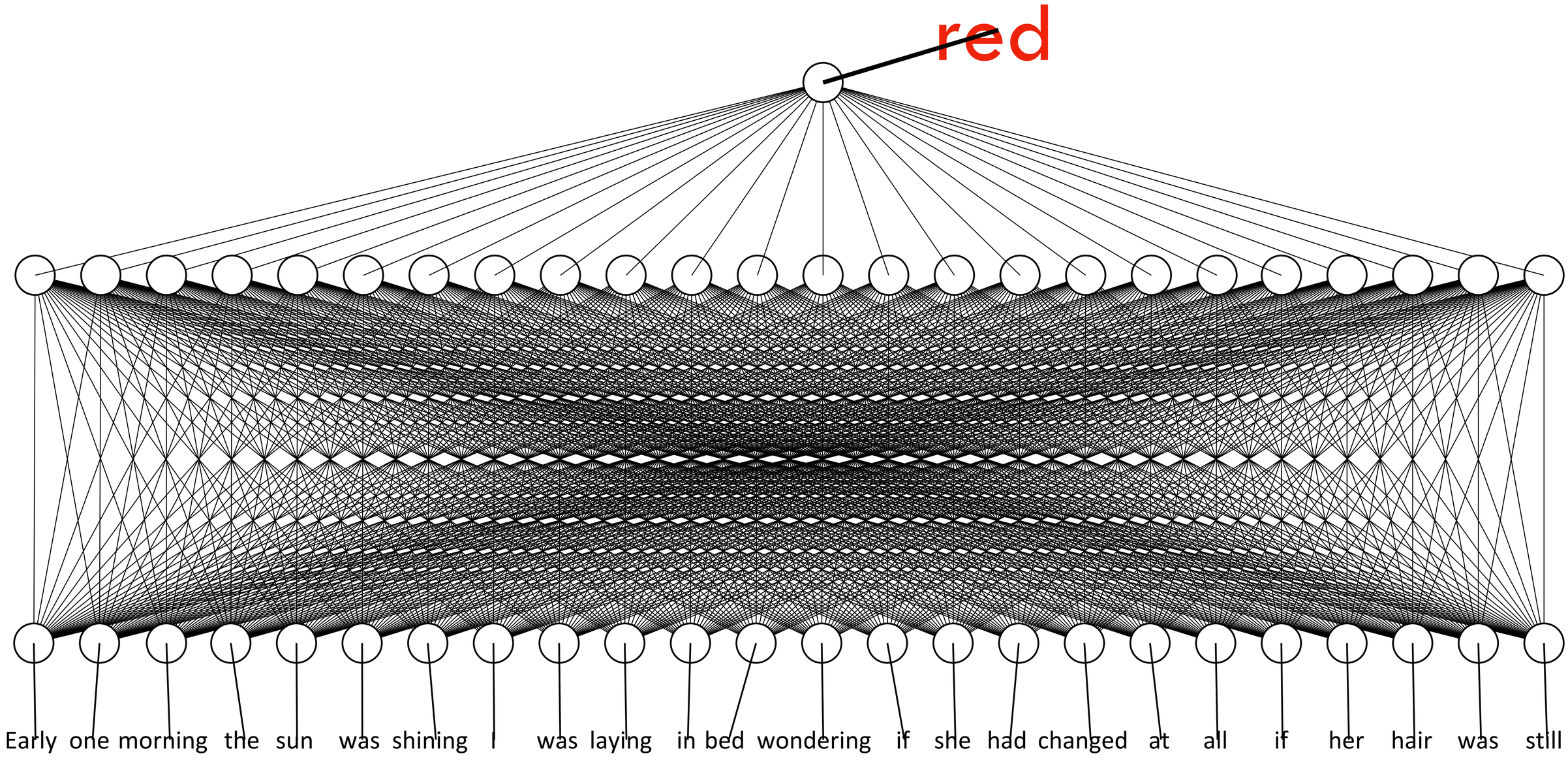
Fourier Series: $f(x) =$  $+ \ldots$

Taylor Series: $f(x) =$  $+ \ldots$

Neural Network:




red
neural network
Early one morning the sun was shining I was laying in bed wondering if she had changed at all if her hair was still

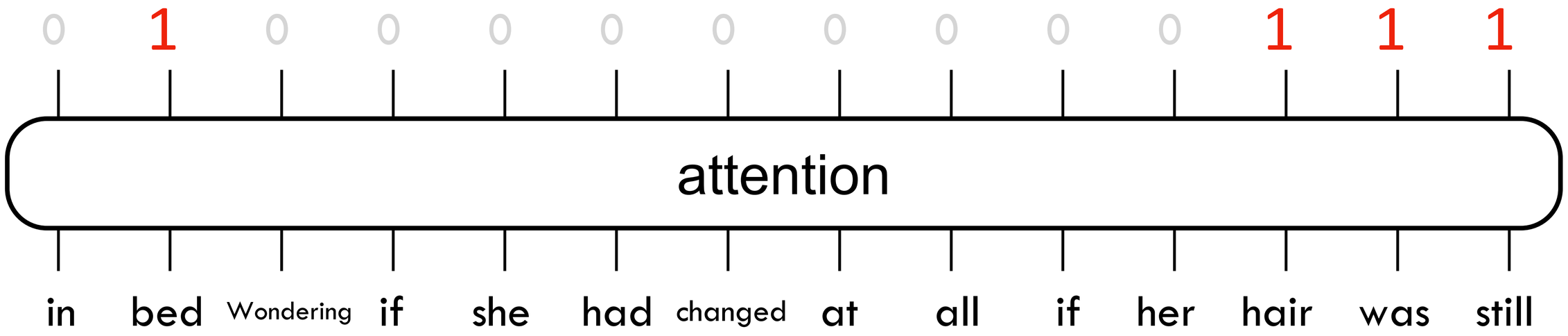

red
Early one morning the sun was shining I was laying in bed wondering if she had changed at all if her hair was still

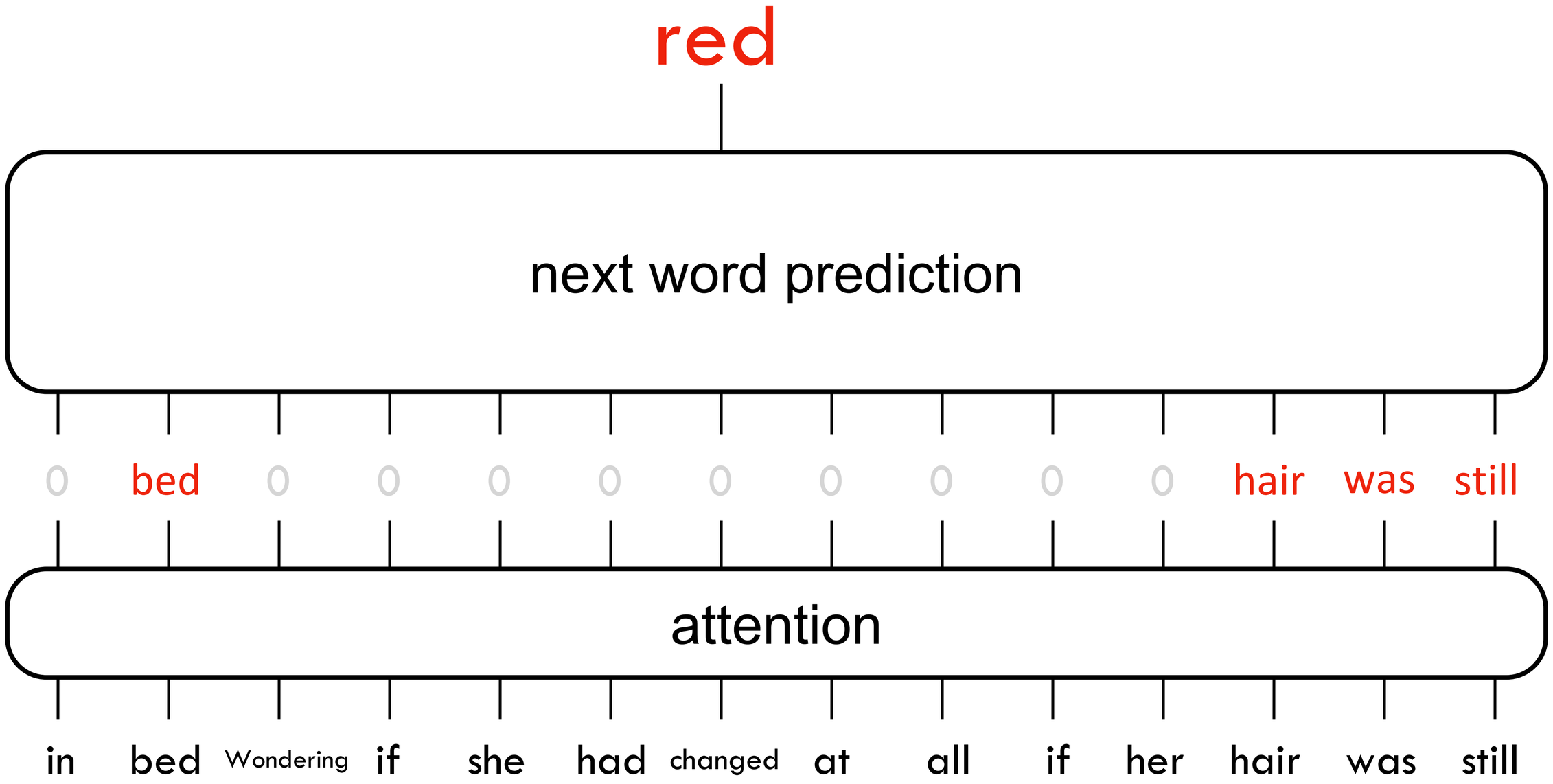
Early one morning the sun was shining I was laying in bed

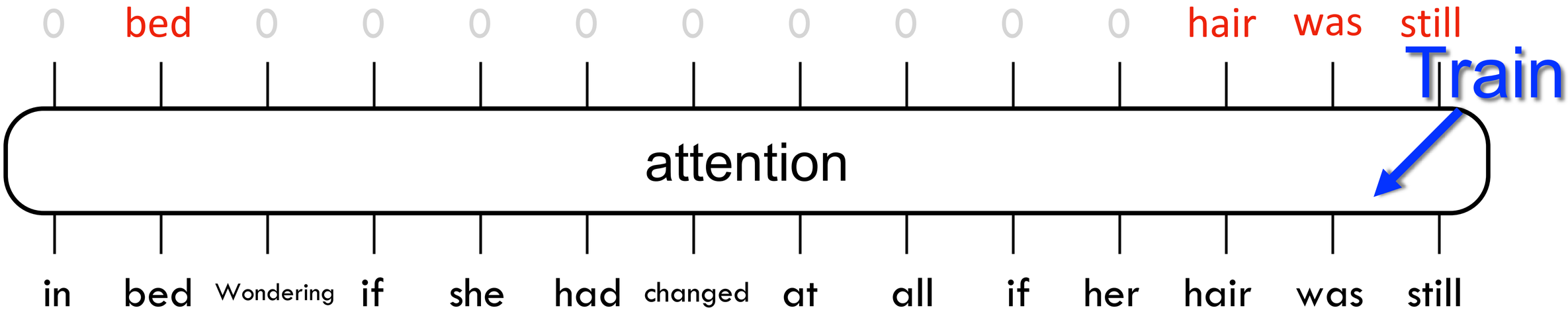
Wondering if she had changed at all if her hair was still ?

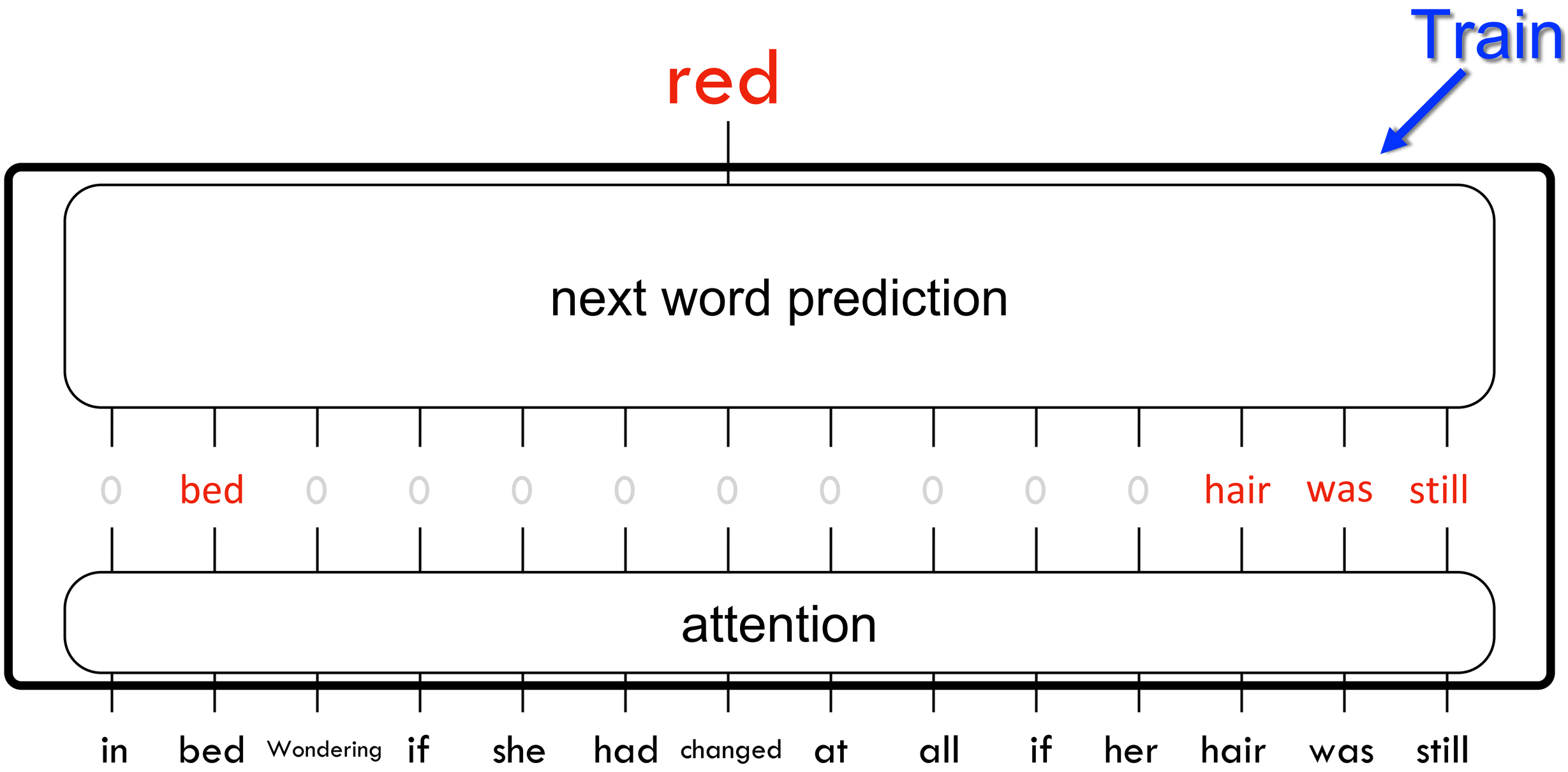
bed

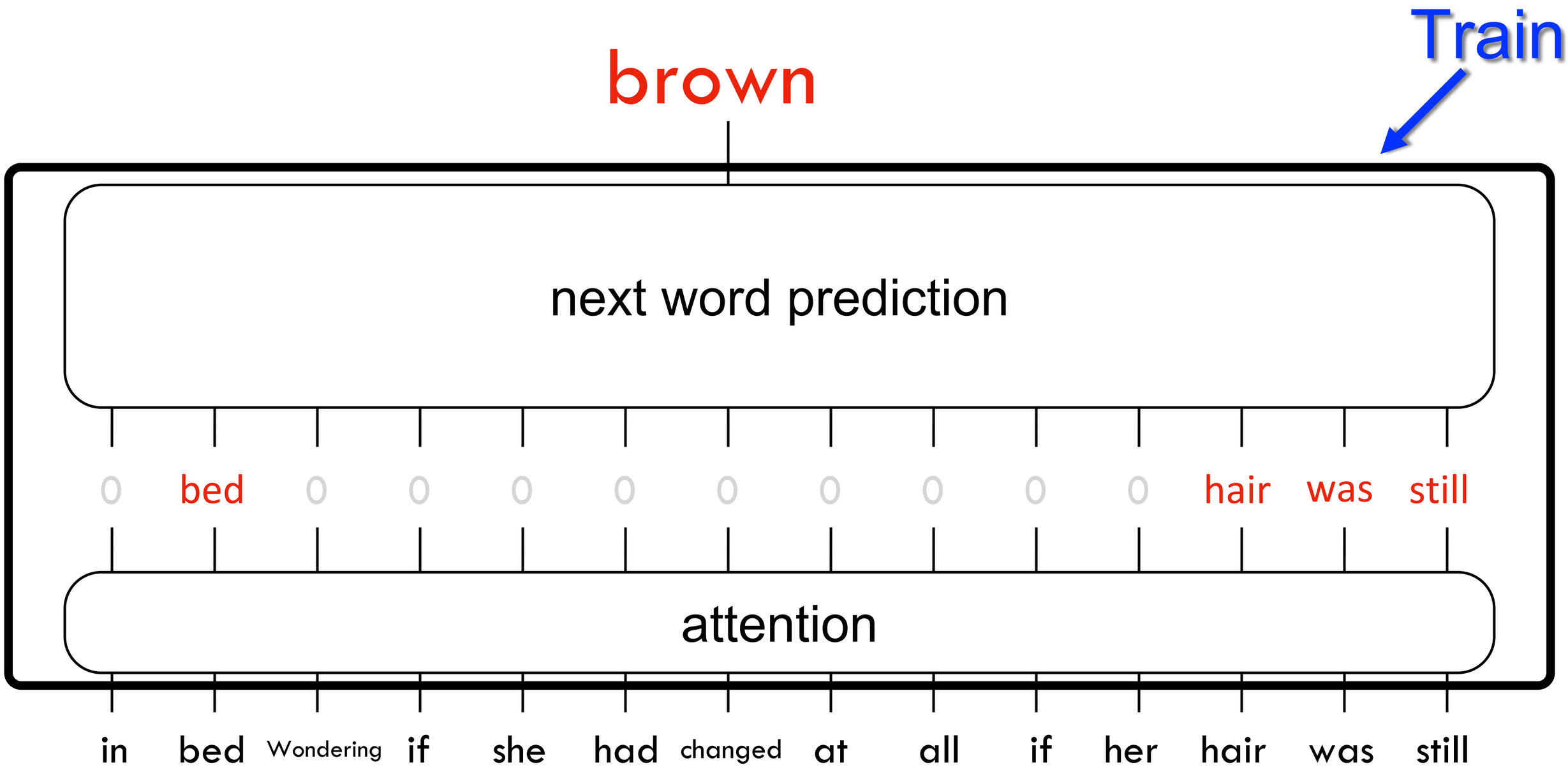
hair was still red

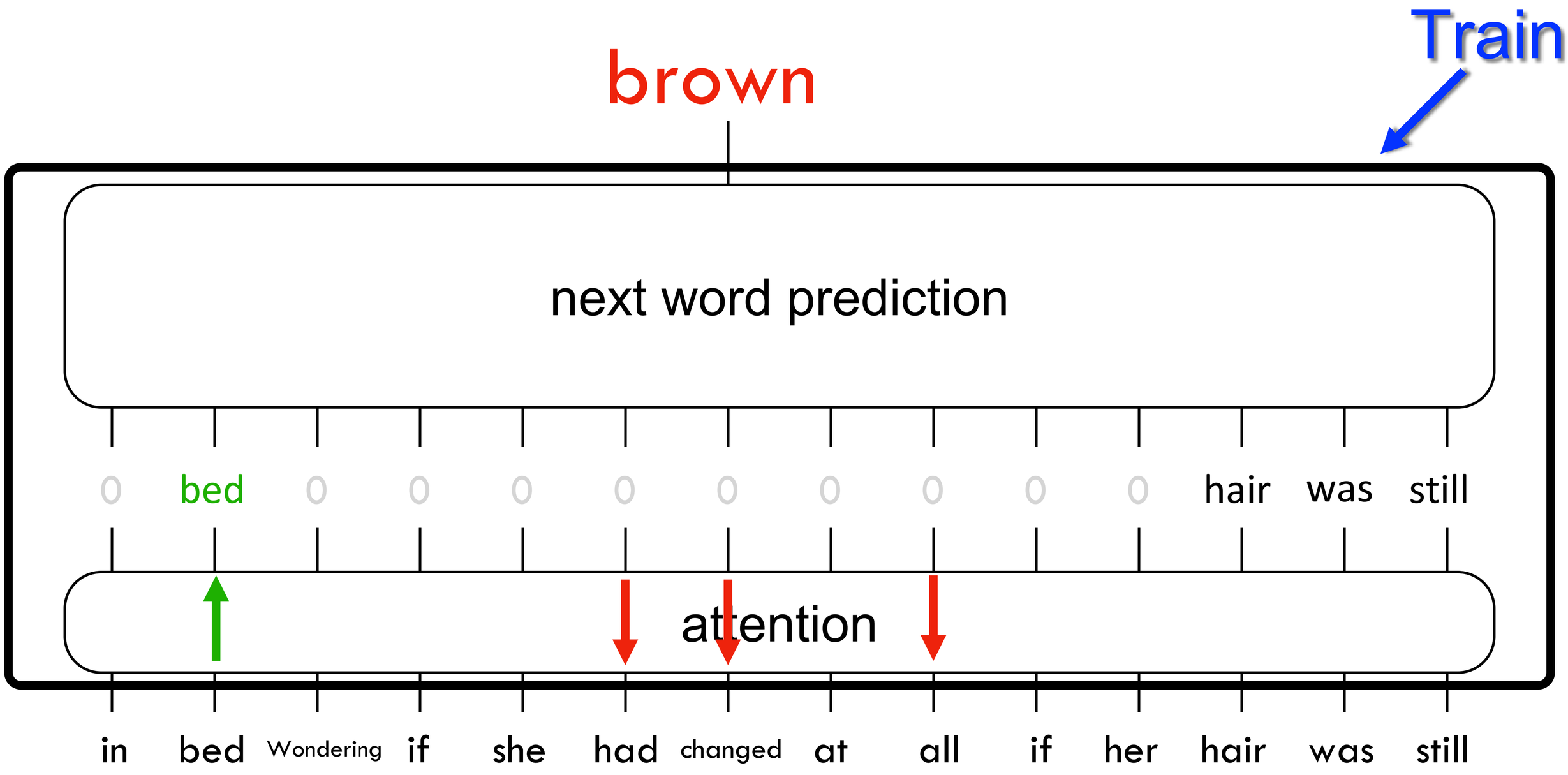

0 1 0 0 0 0 0 0 0 0 0 1 1 1
attention
in bed Wondering if she had changed at all if her hair was still

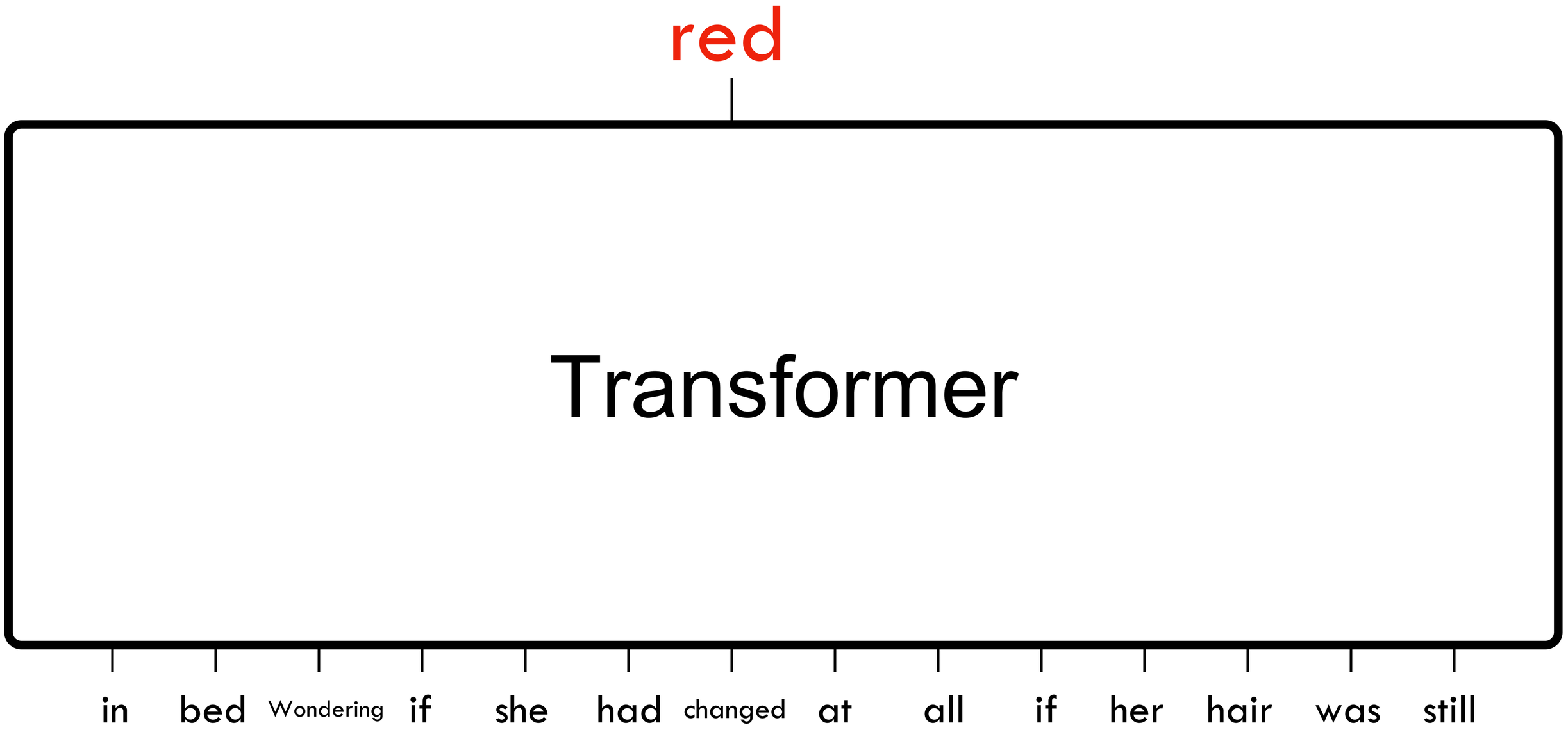

red
next word prediction
0 bed 0 0 0 0 0 0 0 0 0 hair was still
attention
in bed Wondering if she had changed at all if her hair was still

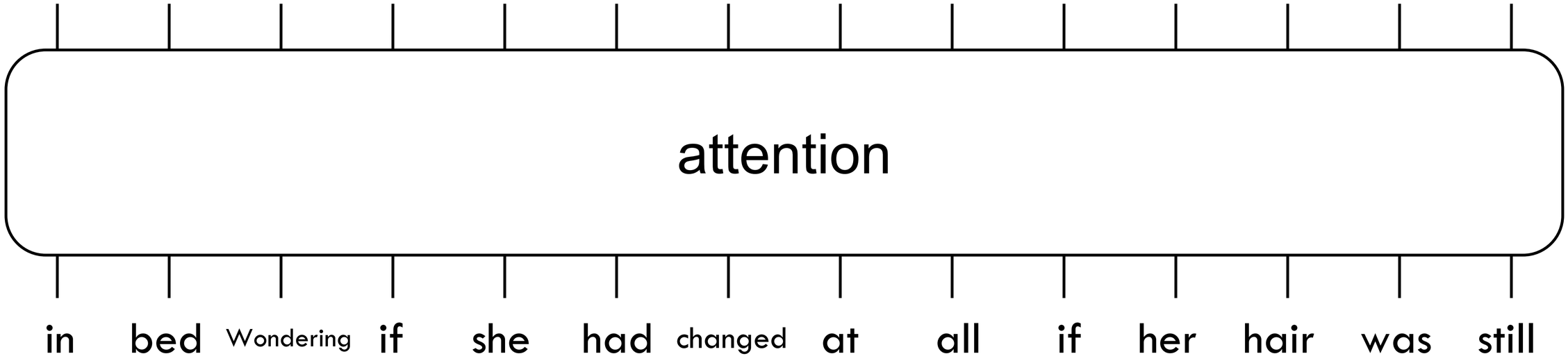

0 bed 0 0 0 0 0 0 0 0 0 hair was still
Train
attention
in bed Wondering if she had changed at all if her hair was still

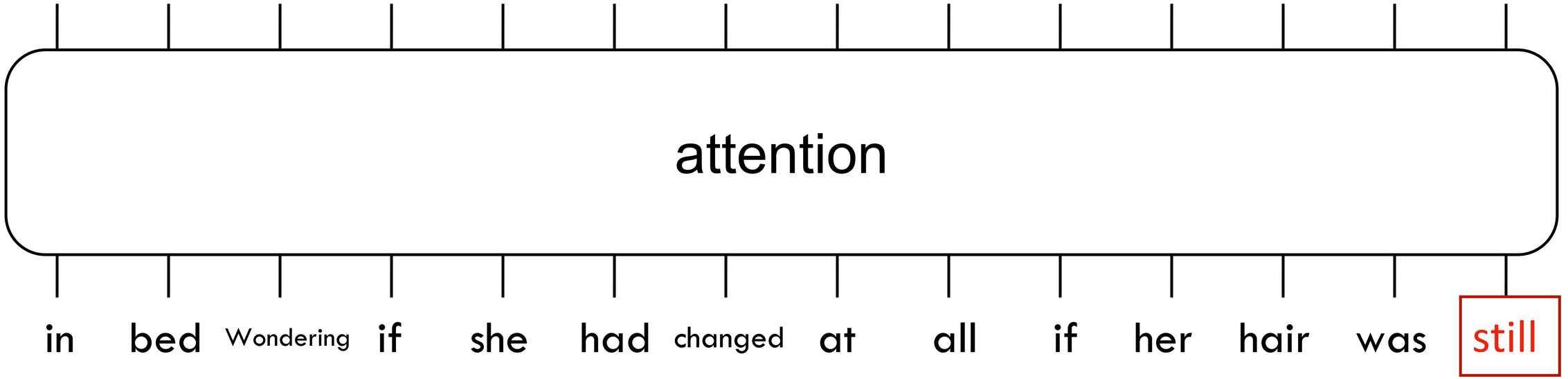

Train
red
next word prediction
0 bed 0 0 0 0 0 0 0 0 0 hair was still
attention
in bed Wondering if she had changed at all if her hair was still

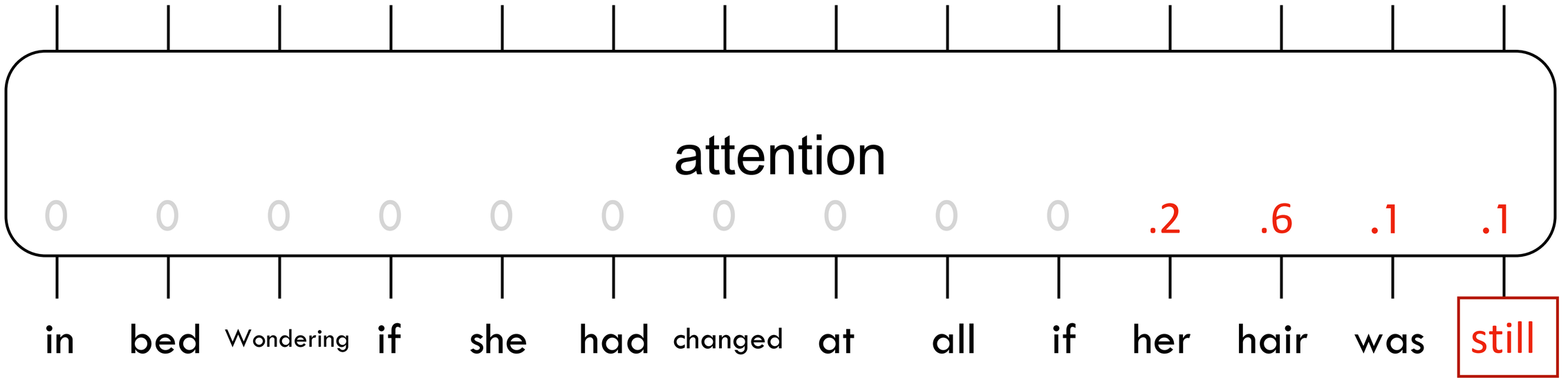

Train
brown
next word prediction
0 bed 0 0 0 0 0 0 0 0 0 hair was still
attention
in bed Wondering if she had changed at all if her hair was still

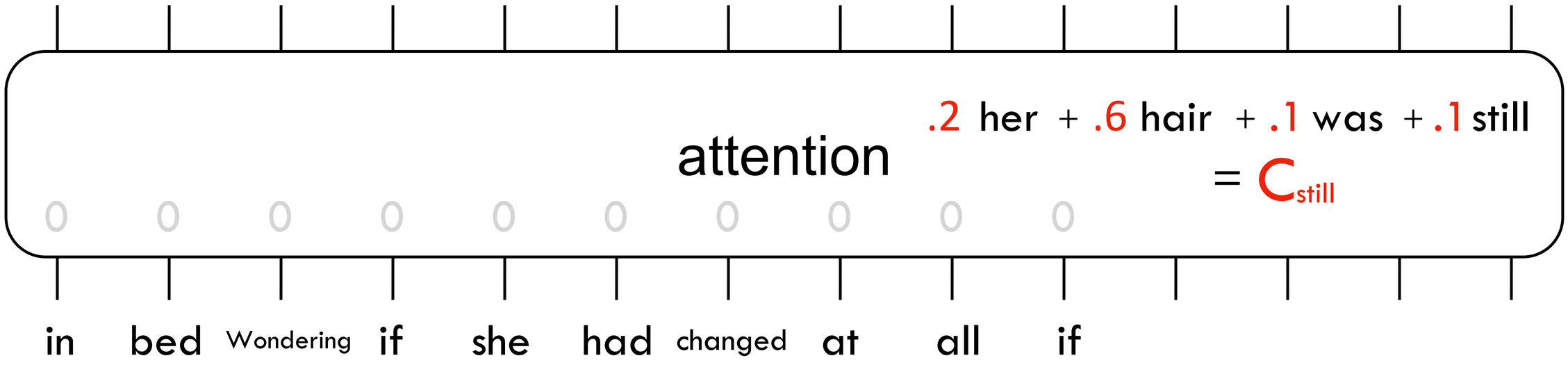

brown
Train
next word prediction
0 bed 0 0 0 0 0 0 0 0 0 hair was still
attention
in bed Wondering if she had changed at all if her hair was still

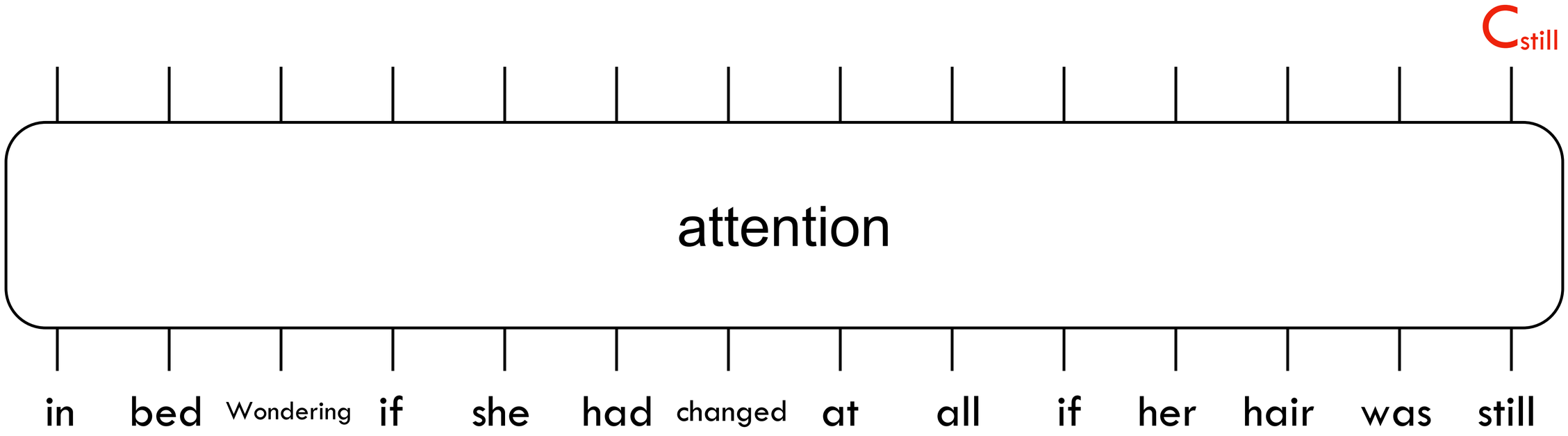

red
Transformer
in
bed
Wondering
if
she
had
changed
at
all
if
her
hair
was
still

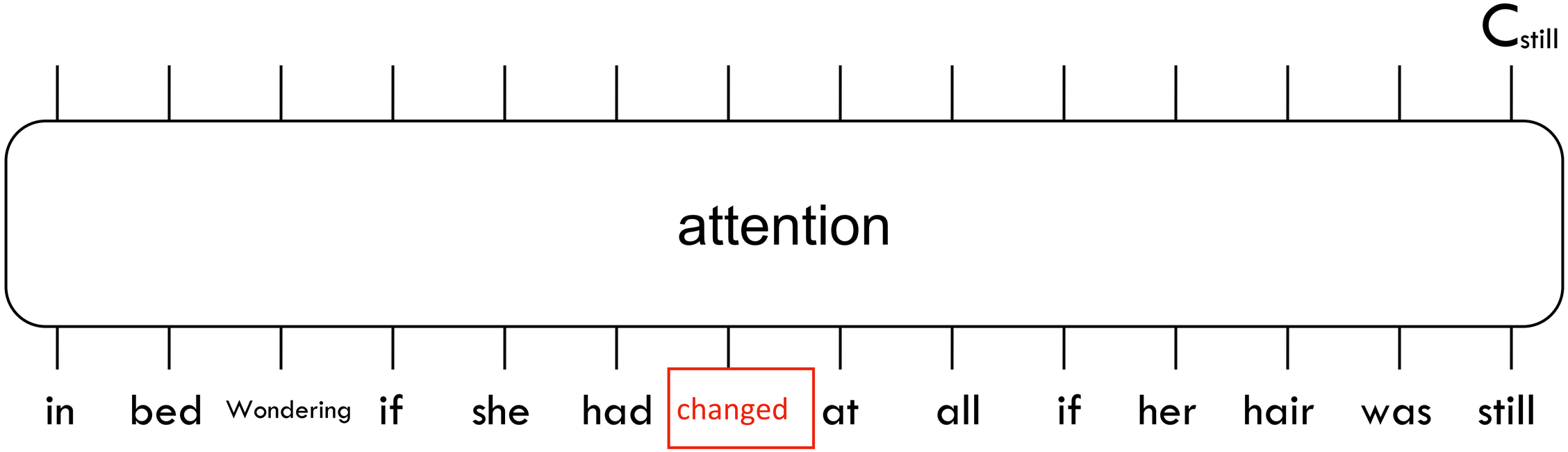

attention
in bed Wondering if she had changed at all if her hair was still

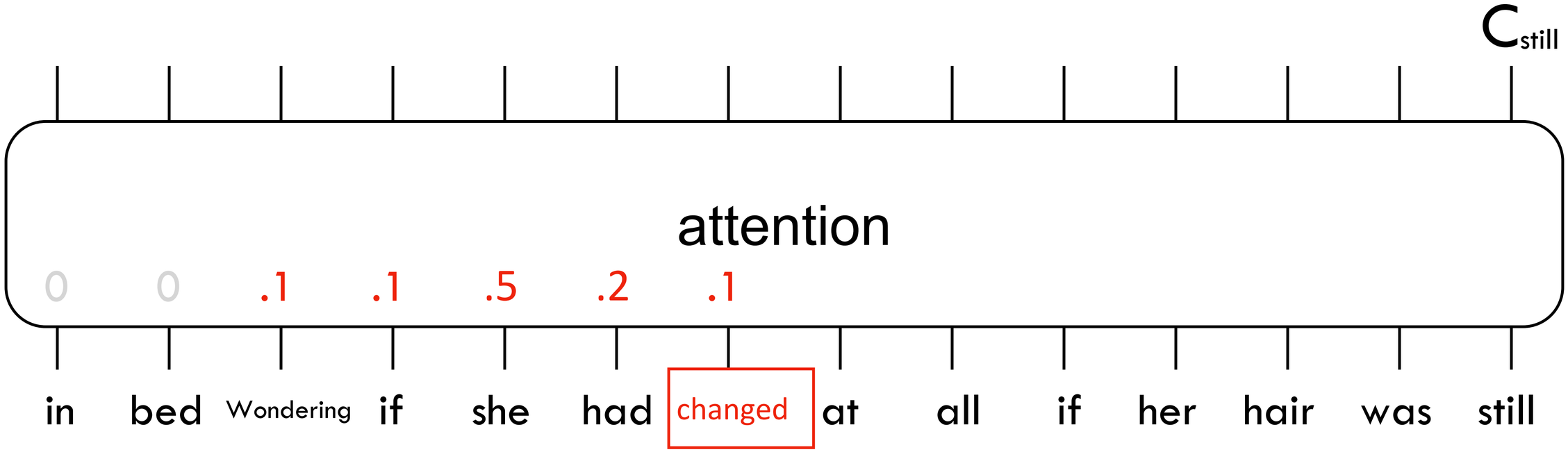

attention
in bed Wondering if she had changed at all if her hair was still

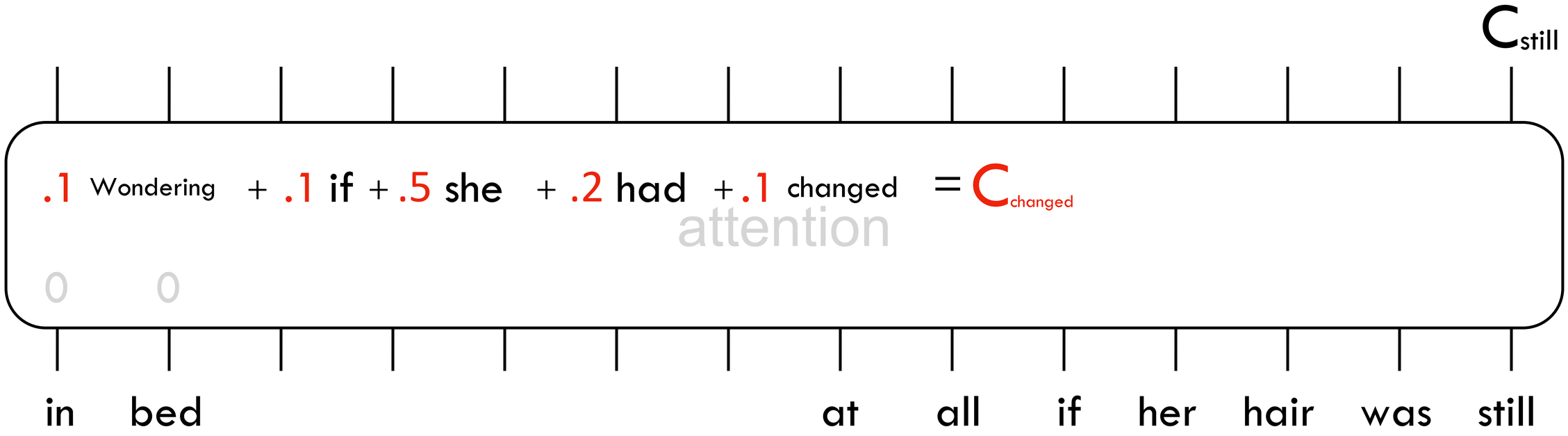

attention
0 0 0 0 0 0 0 0 0 0 .2 .6 .1 .1
in bed Wondering if she had changed at all if her hair was still

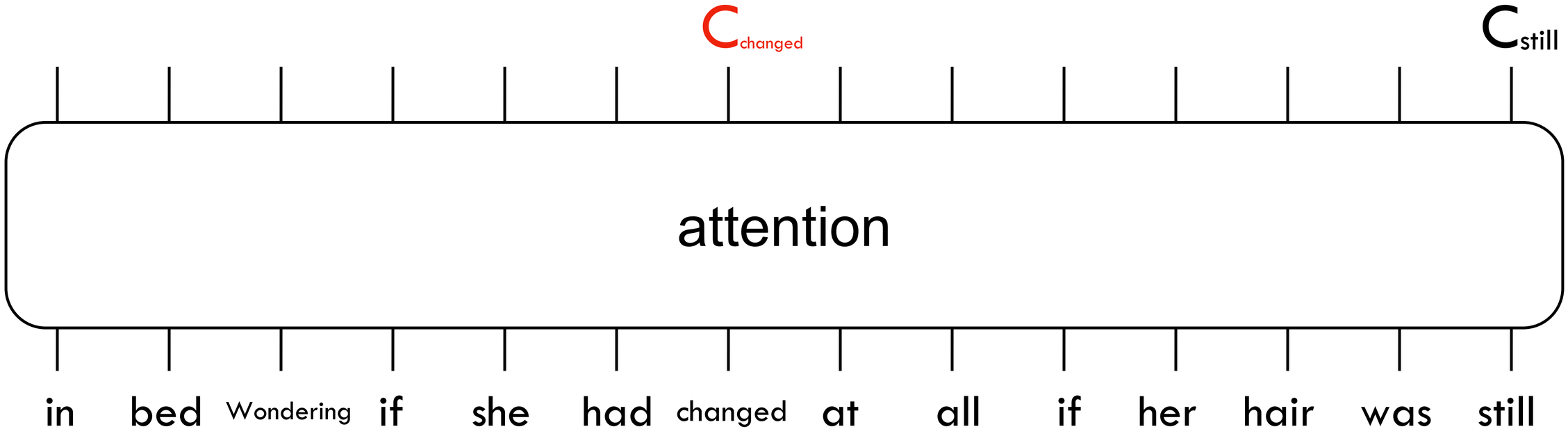

attention
.2 her + .6 hair + .1 was + .1 still
= C still
0 0 0 0 0 0 0 0 0 0
in bed Wondering if she had changed at all if

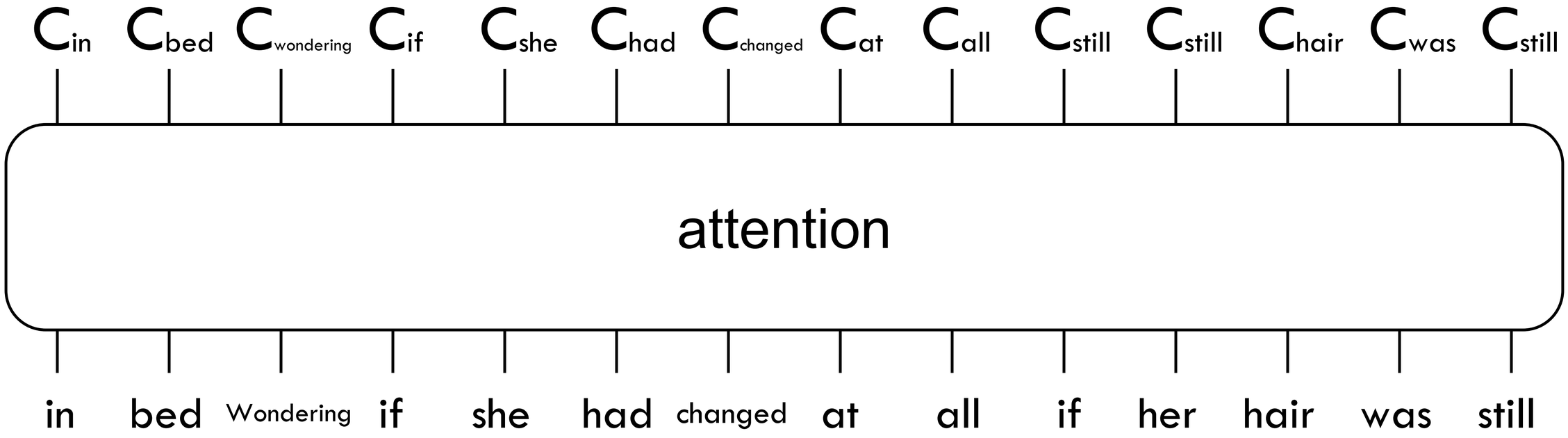

Cstill
attention
in bed Wondering if she had changed at all if her hair was still

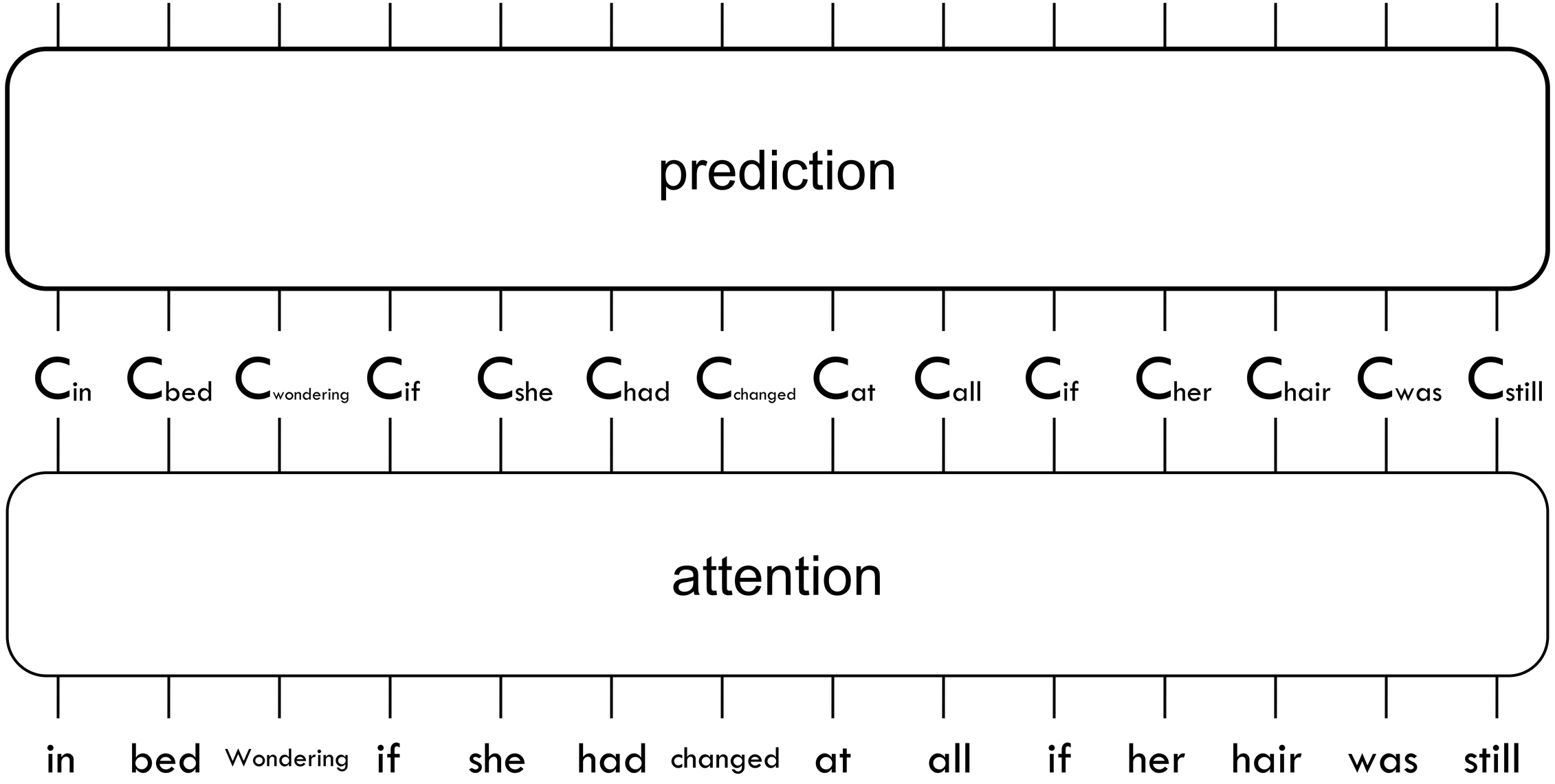

Cstill
attention
in bed Wondering if she had changed at all if her hair was still

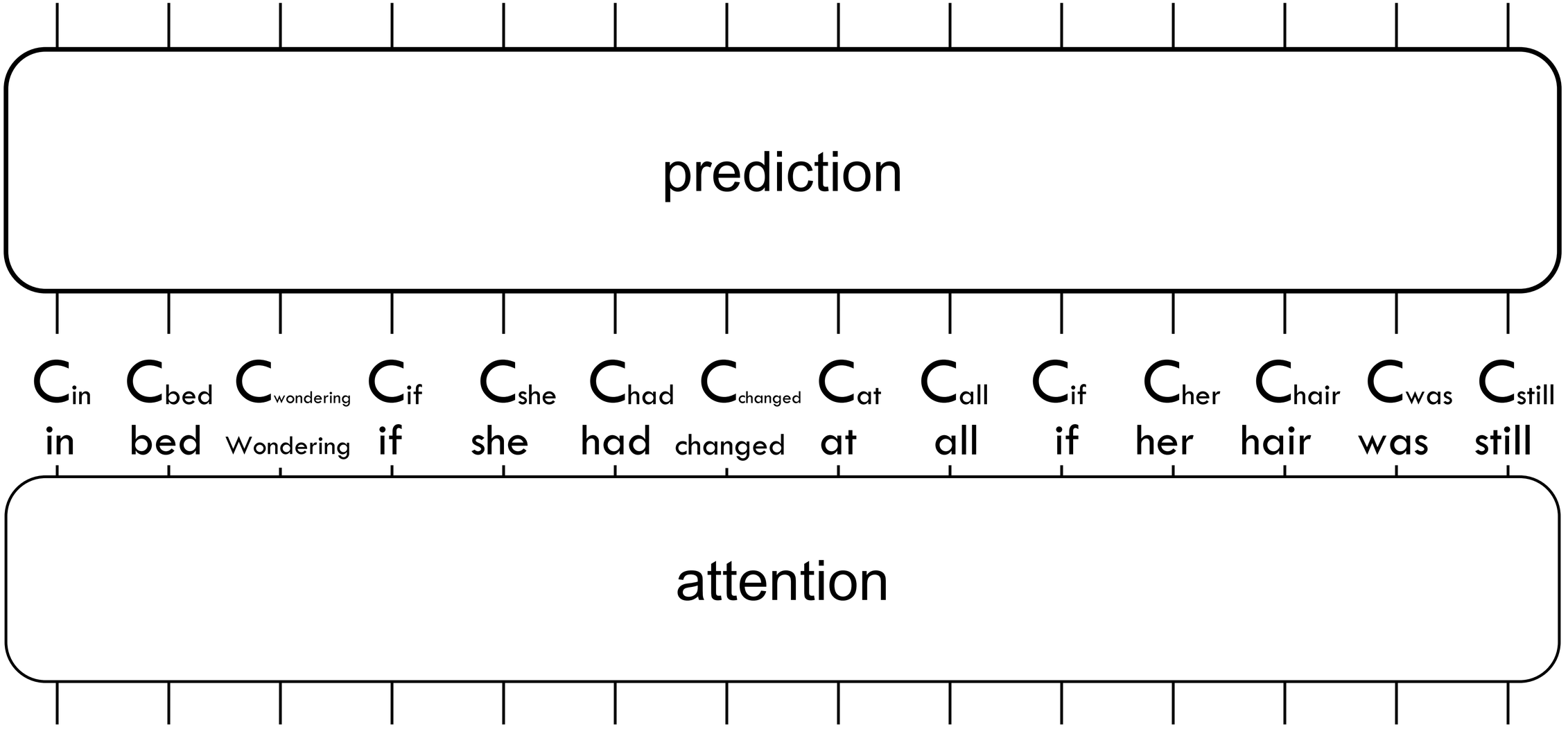

$C_{still}$
attention
0 0 .1 .1 .5 .2 .1
in bed Wondering if she had changed at all if her hair was still

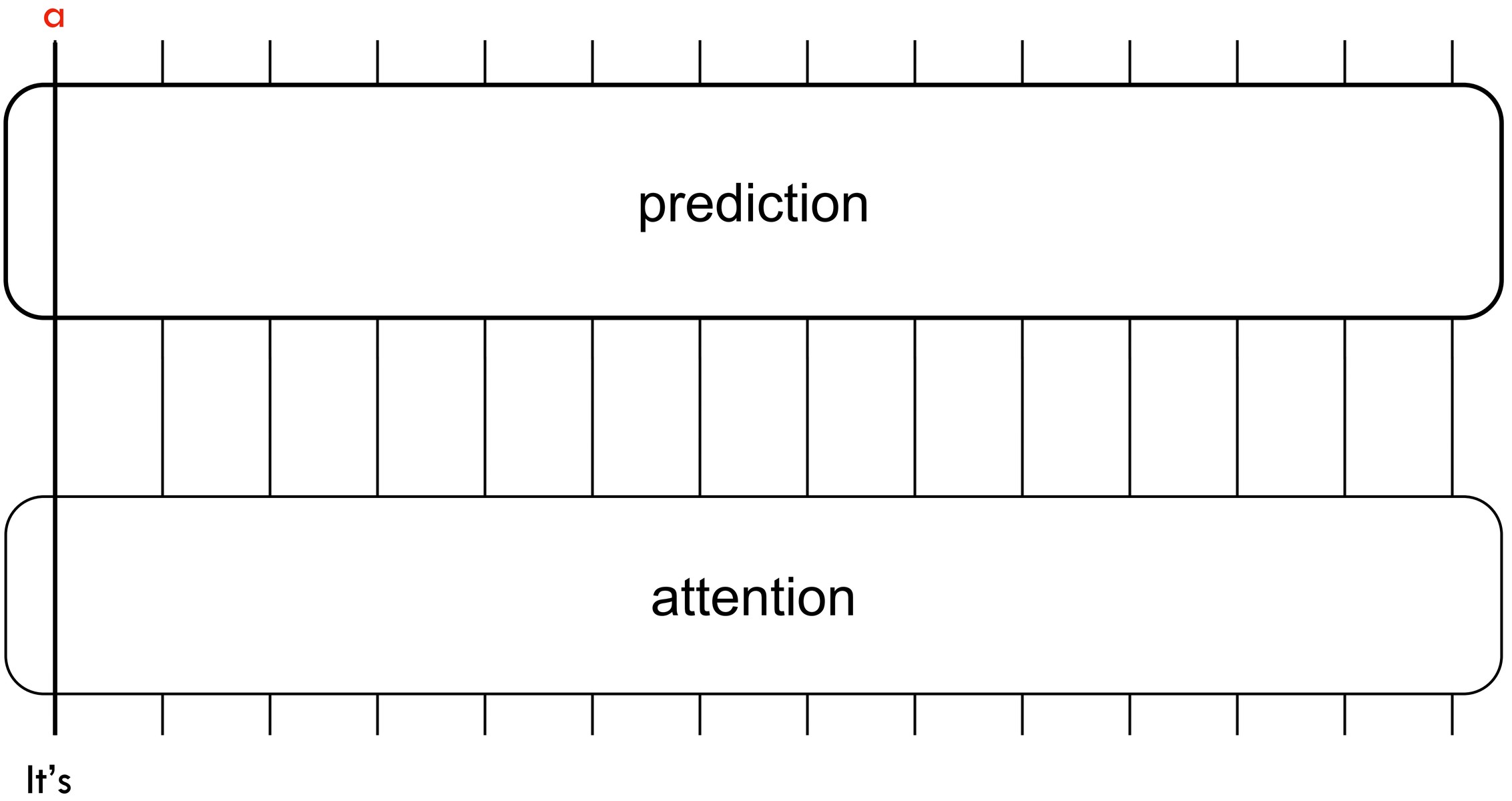

$C_{still}$
.1 Wondering + .1 if + .5 she + .2 had + .1 changed = $C_{changed}$
attention
0 0
in bed at all if her hair was still


Cchanged
Cstill
attention
in bed Wondering if she had changed at all if her hair was still

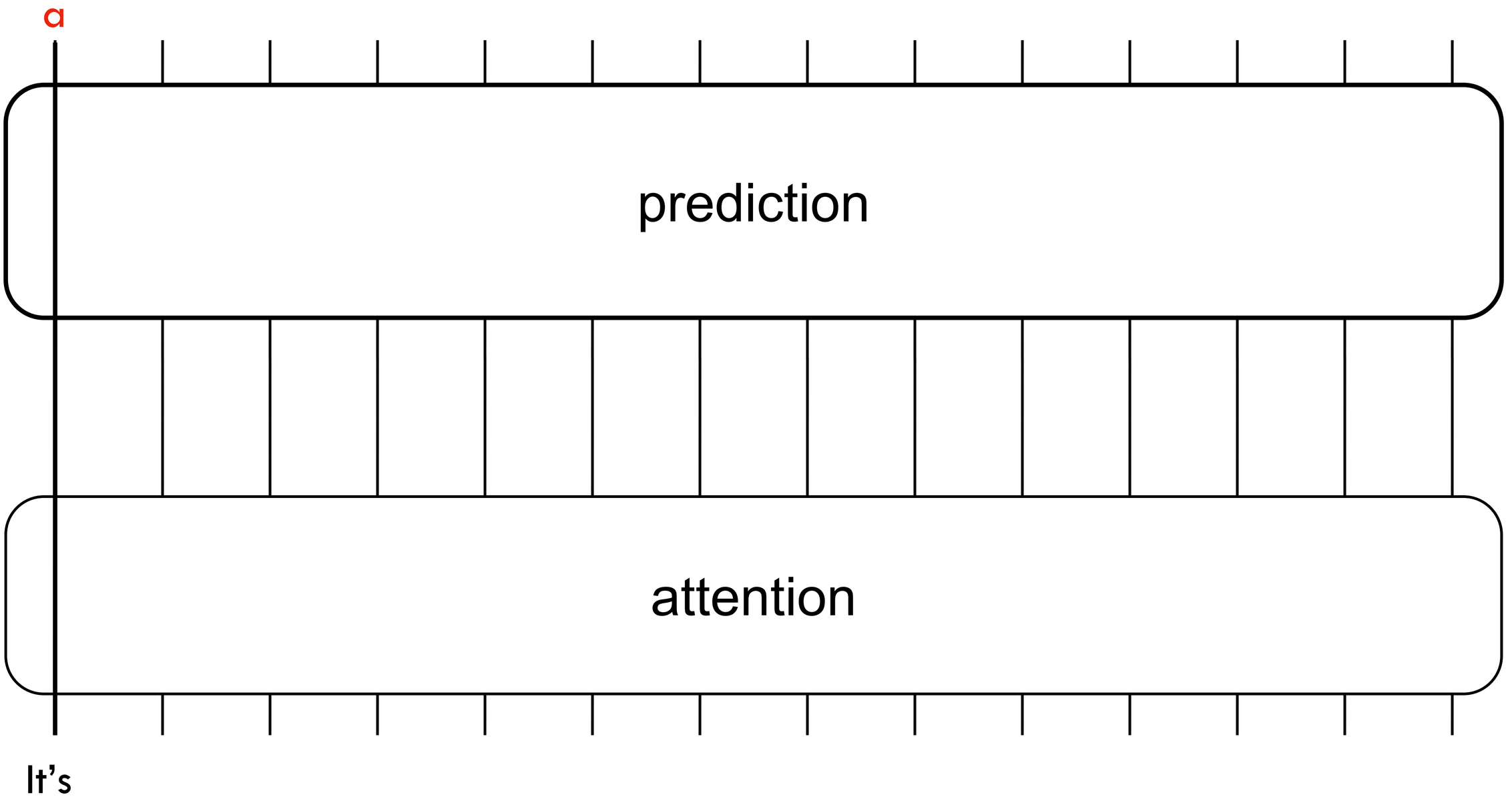

$C_{in}$ $C_{bed}$ $C_{wondering}$ $C_{if}$ $C_{she}$ $C_{had}$ $C_{changed}$ $C_{at}$ $C_{all}$ $C_{still}$ $C_{still}$ $C_{hair}$ $C_{was}$ $C_{still}$
attention
in bed Wondering if she had changed at all if her hair was still

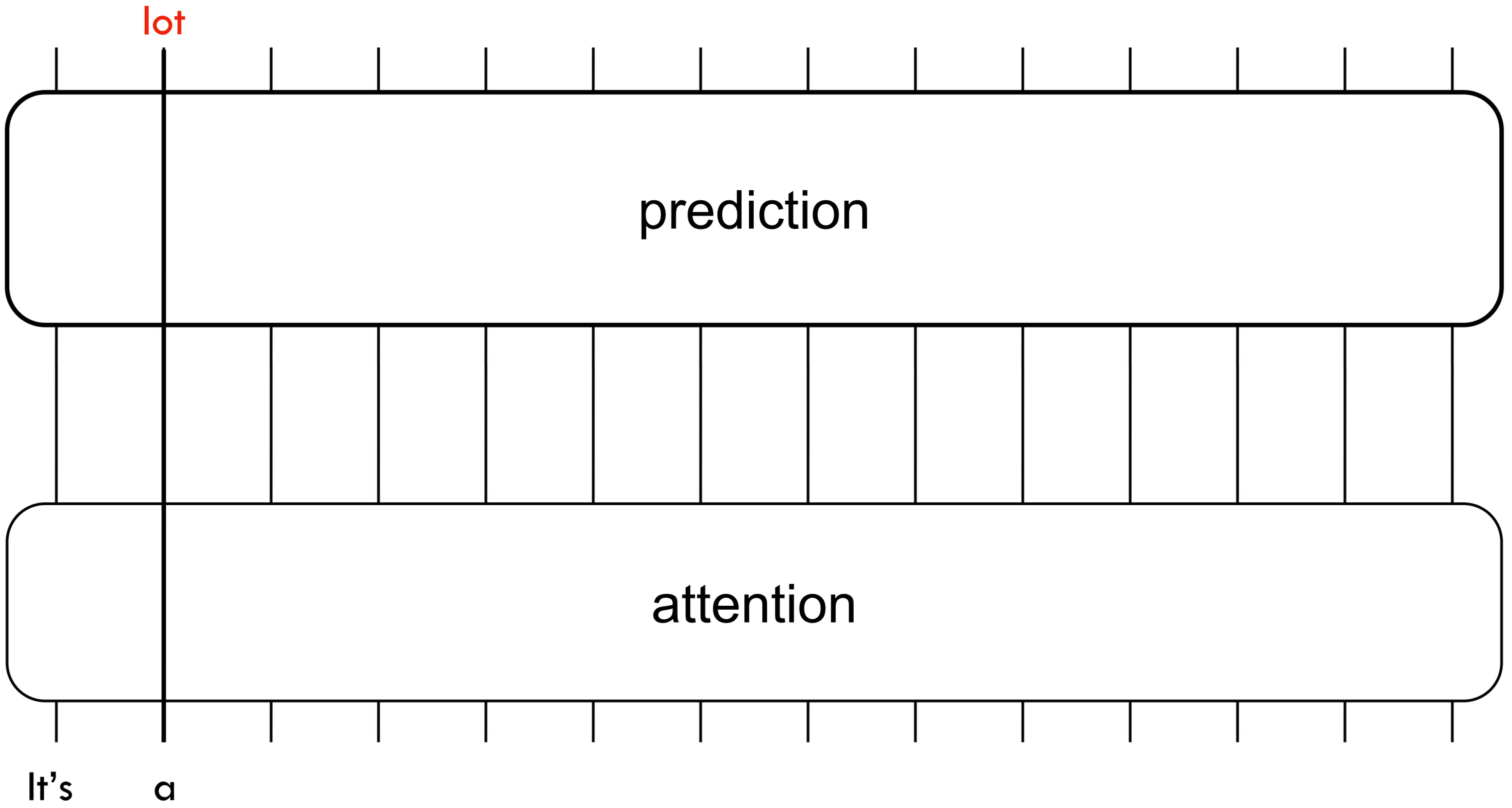

prediction
$C_{in}$ $C_{bed}$ $C_{wondering}$ $C_{if}$ $C_{she}$ $C_{had}$ $C_{changed}$ $C_{at}$ $C_{all}$ $C_{if}$ $C_{her}$ $C_{hair}$ $C_{was}$ $C_{still}$
attention
in bed Wondering if she had changed at all if her hair was still

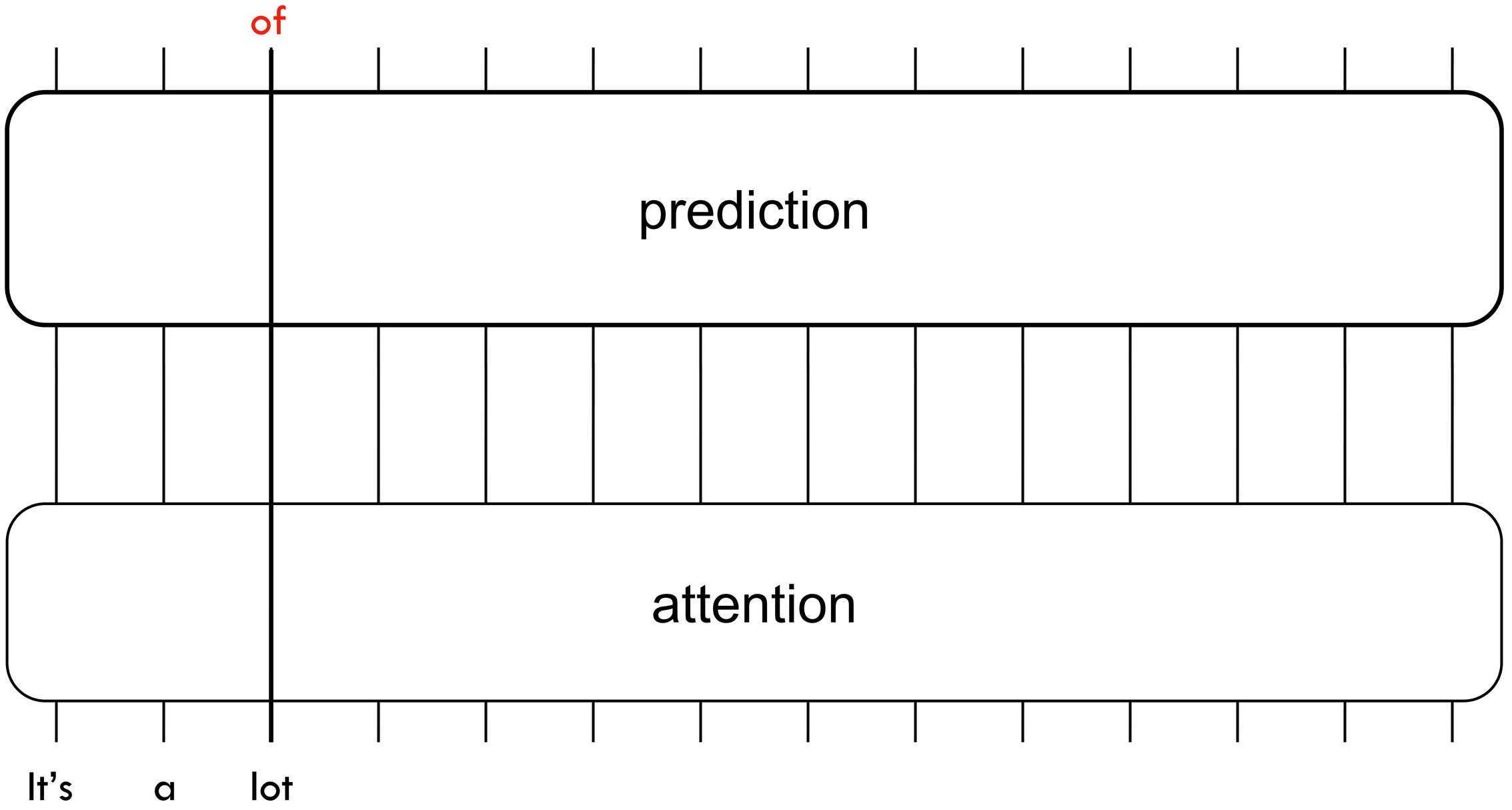

prediction
$C_{in}$ $C_{bed}$ $C_{wondering}$ $C_{if}$ $C_{she}$ $C_{had}$ $C_{changed}$ $C_{at}$ $C_{all}$ $C_{if}$ $C_{her}$ $C_{hair}$ $C_{was}$ $C_{still}$
in bed Wondering if she had changed at all if her hair was still
attention

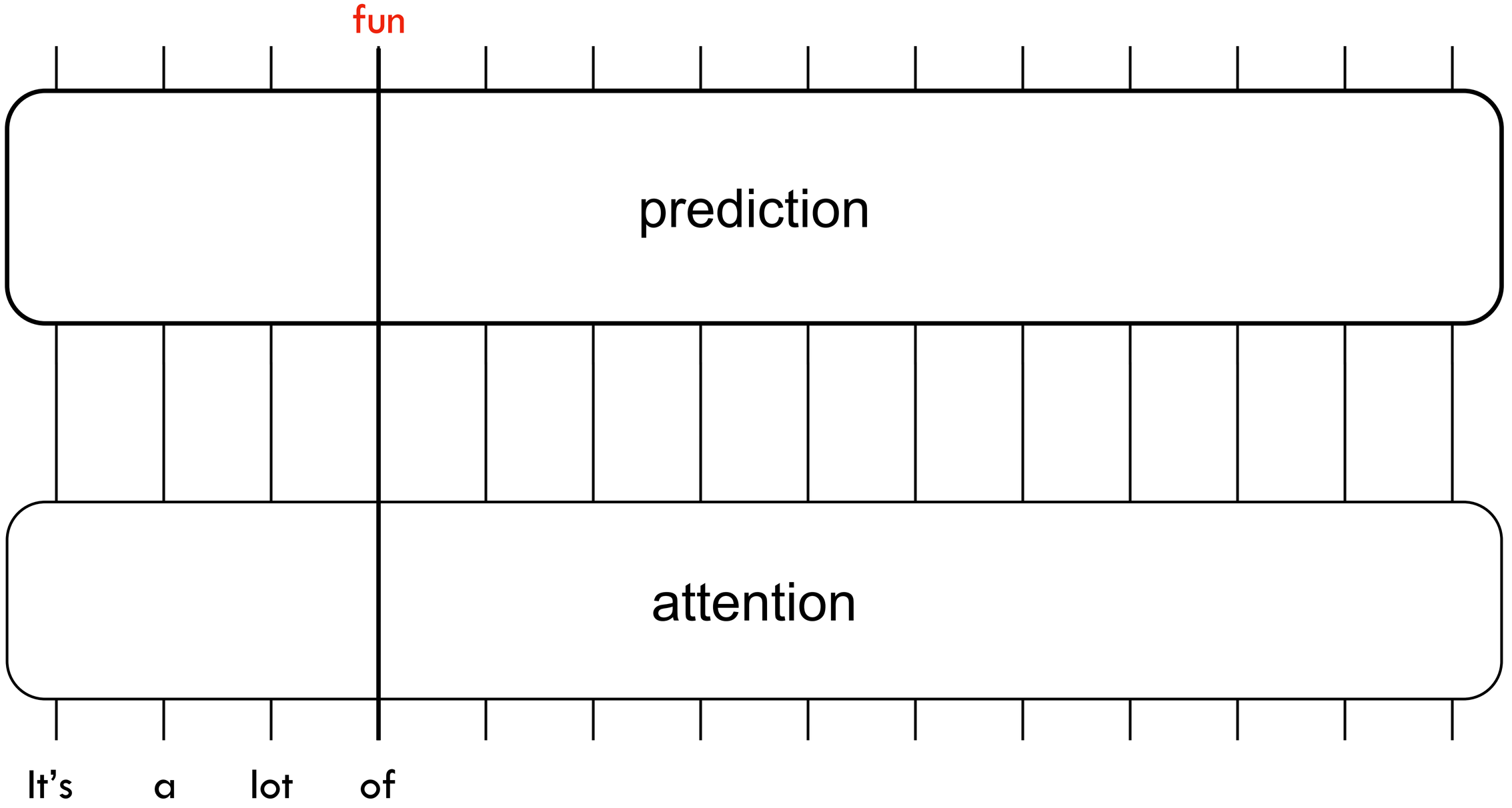

a
prediction
attention
It's

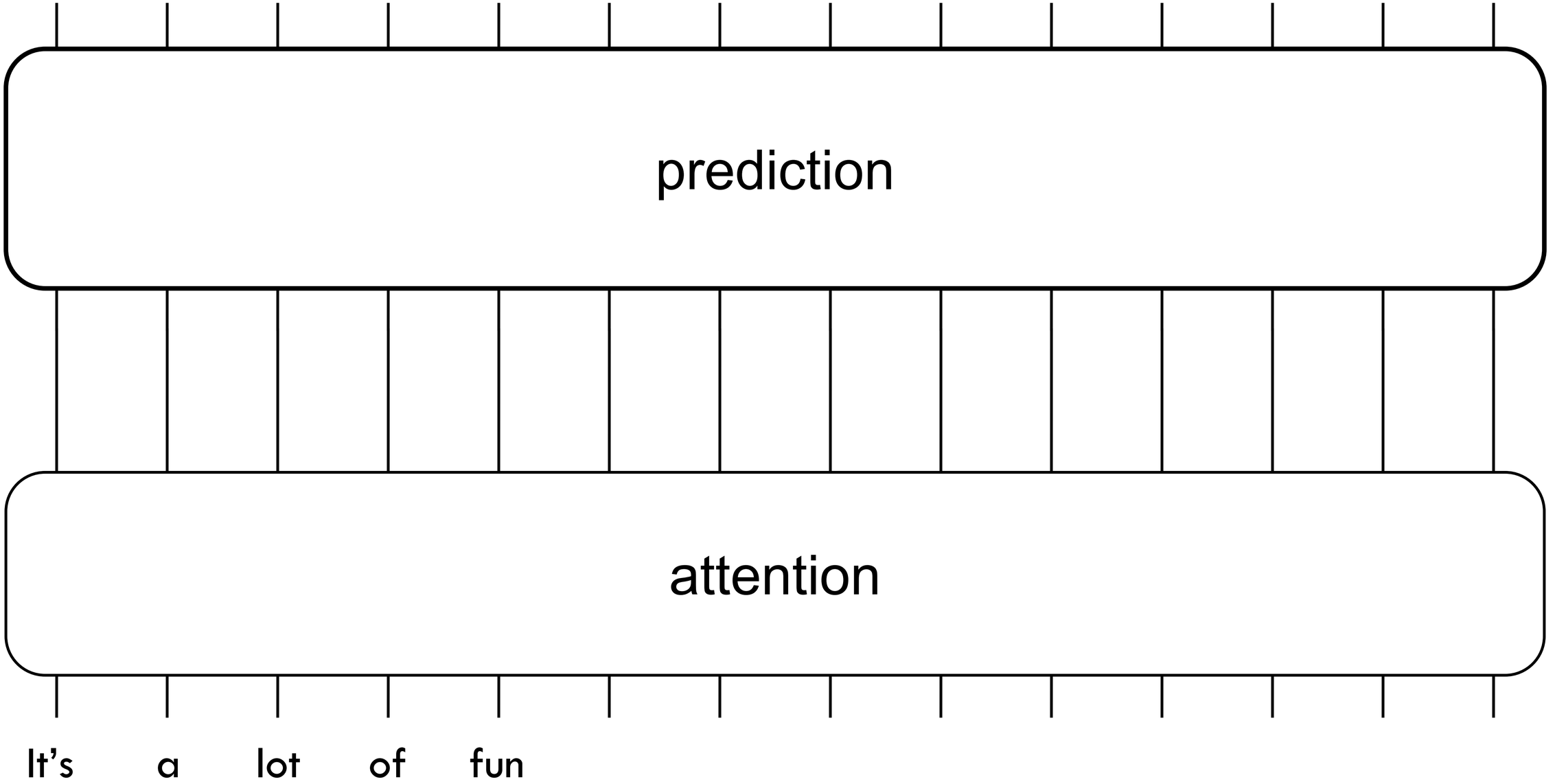
| a | the | looking | possible | getting |
| --- | --- | --- | --- | --- |
| 0.4 | 0.3 | 0.1 | 0.1 | 0.1 |

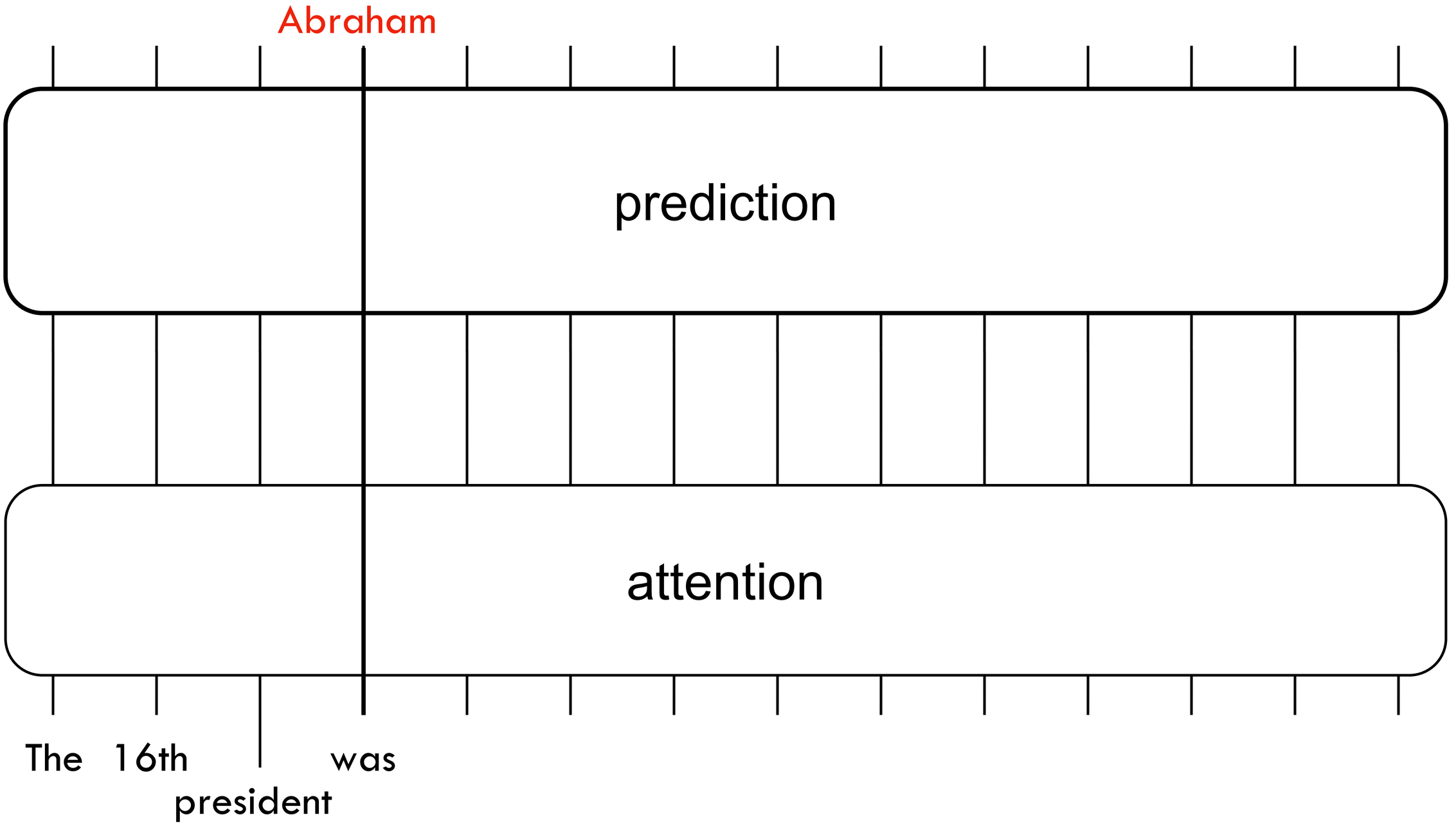

a
prediction
attention
It's

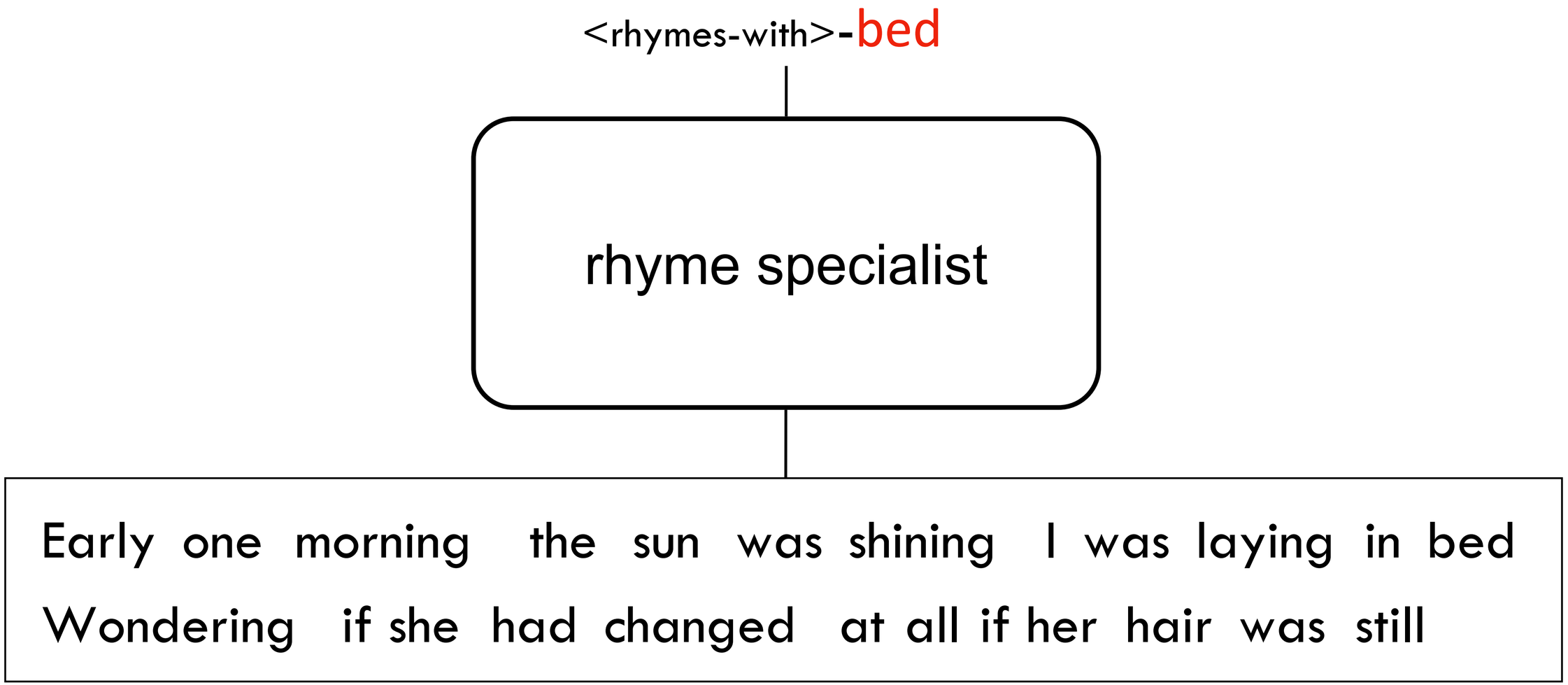

lot
prediction
attention
It's
a

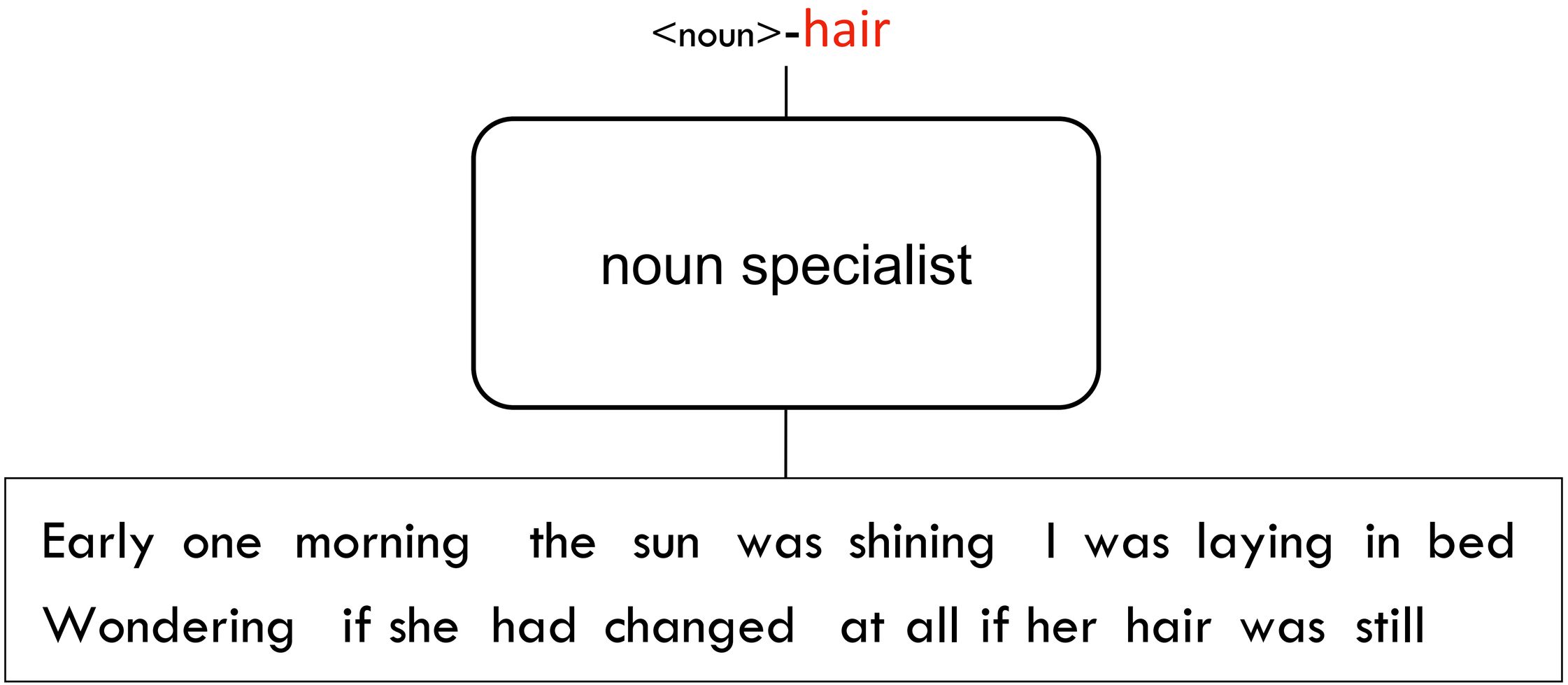

of
prediction
attention
It's
a
lot

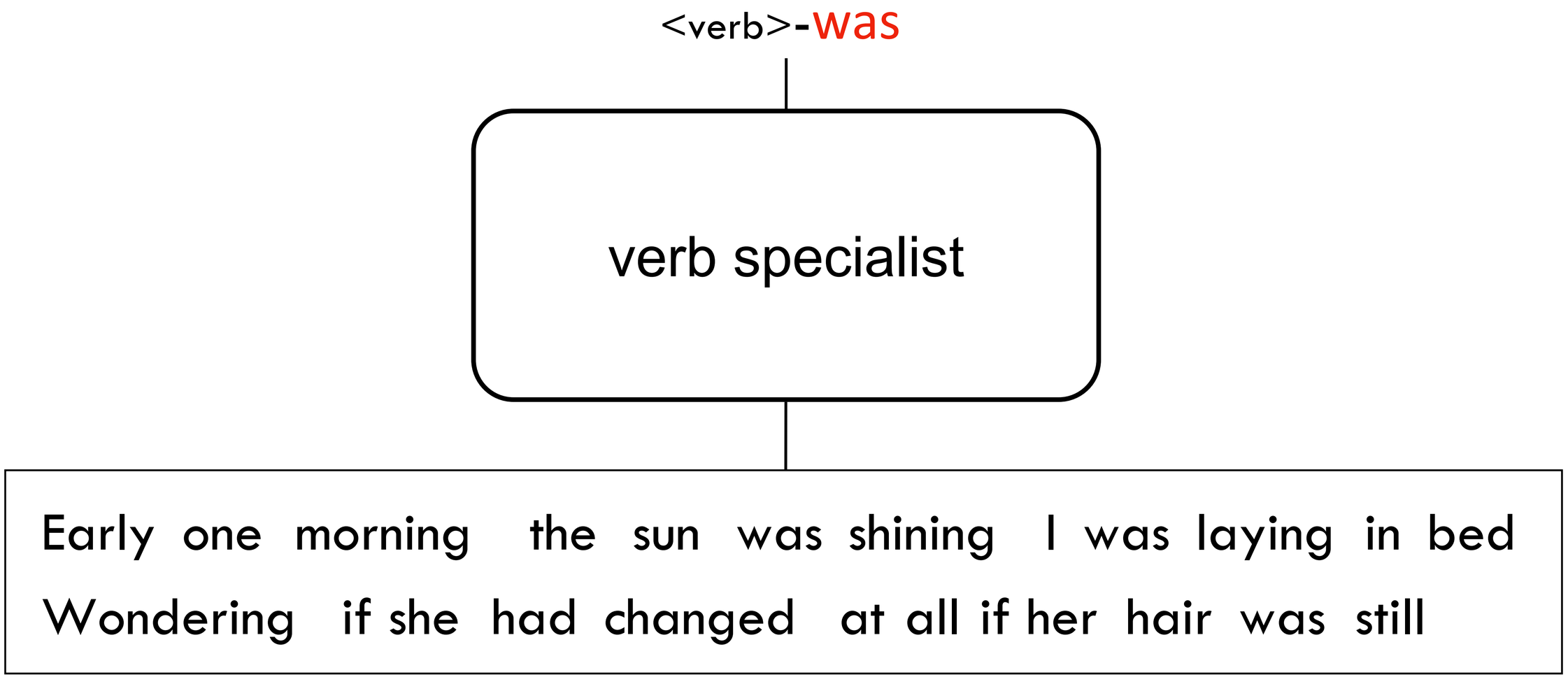

fun
prediction
attention
It's
a
lot
of

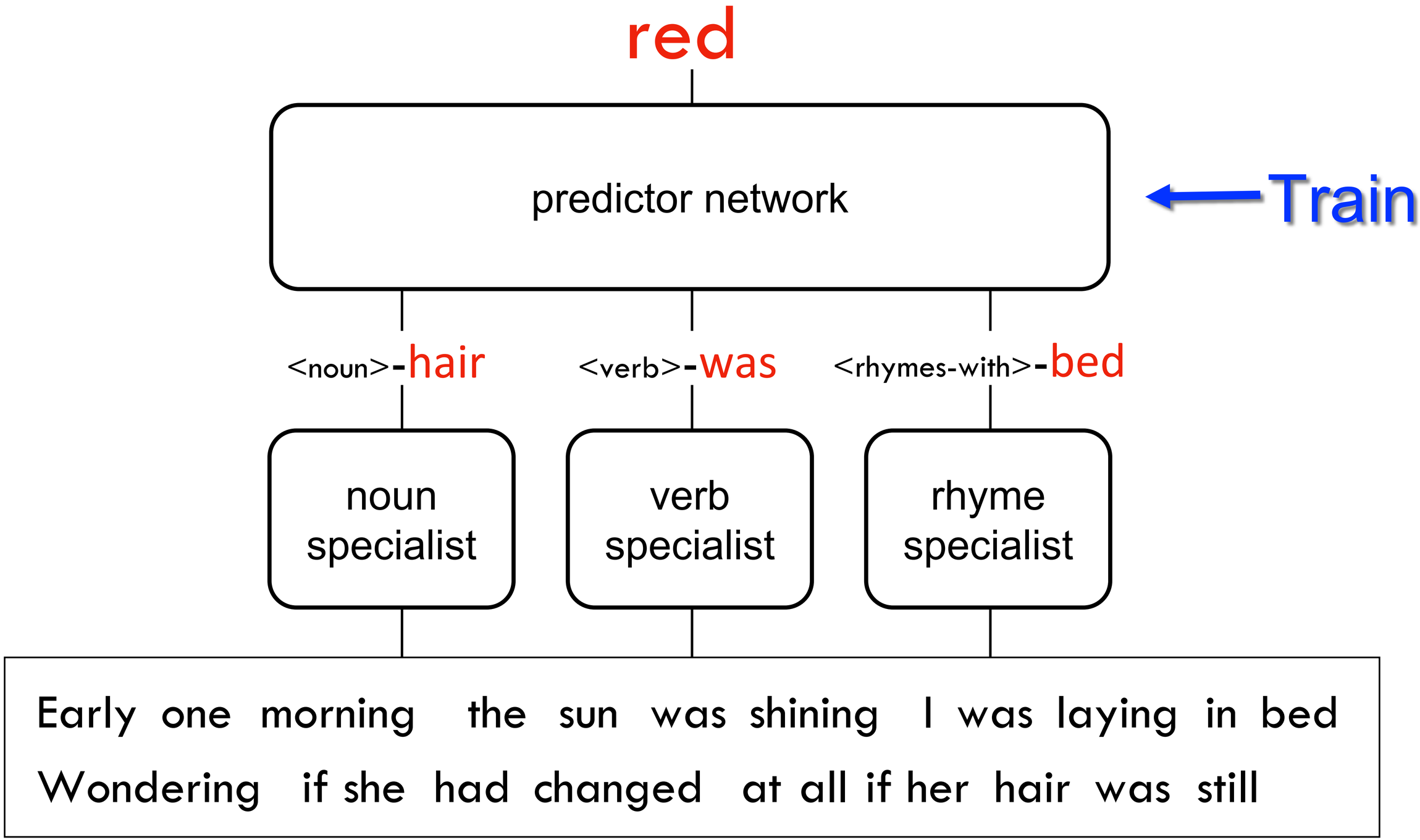

prediction
attention
It's
a
lot
of
fun

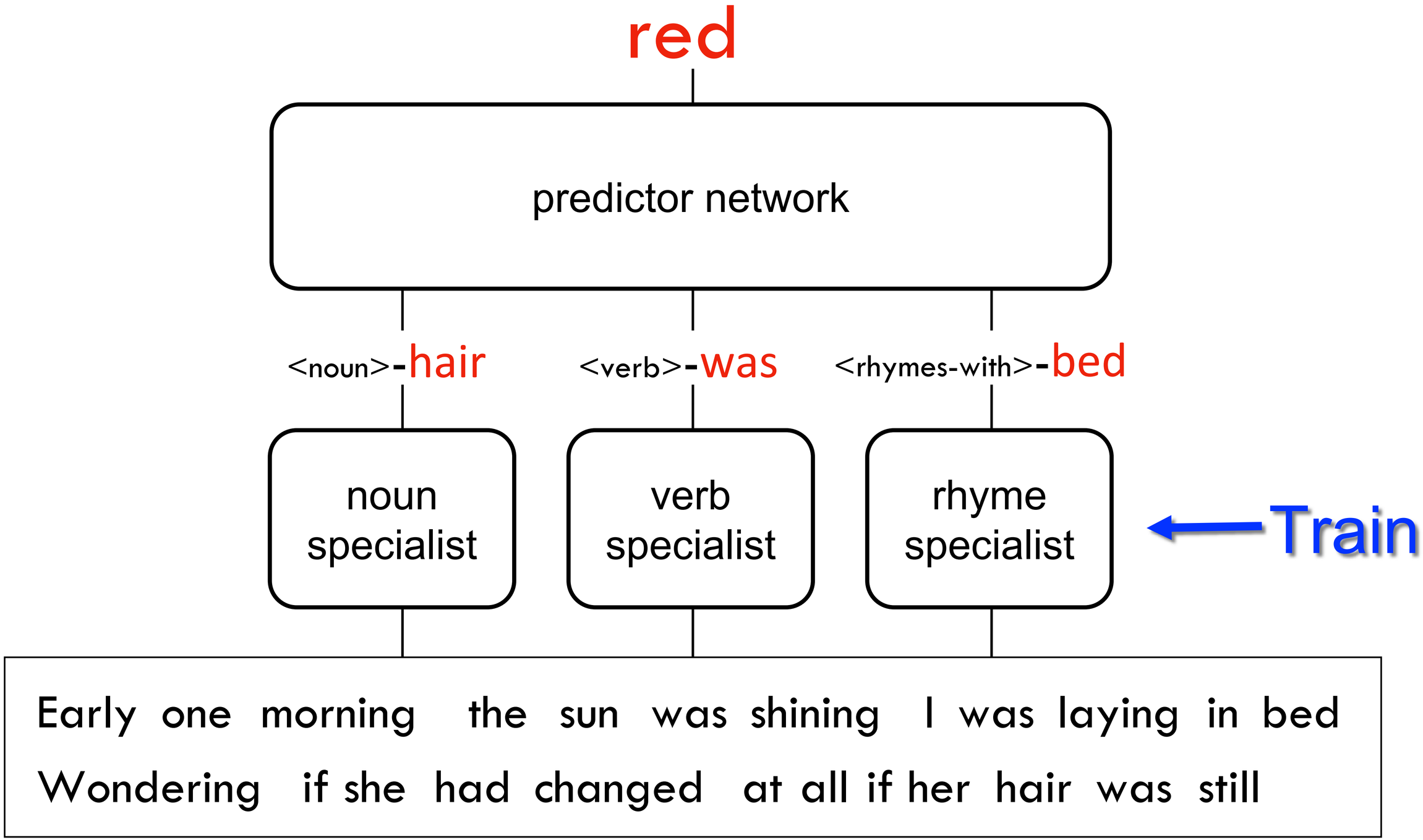

Abraham
prediction
attention
The
16th
president
was

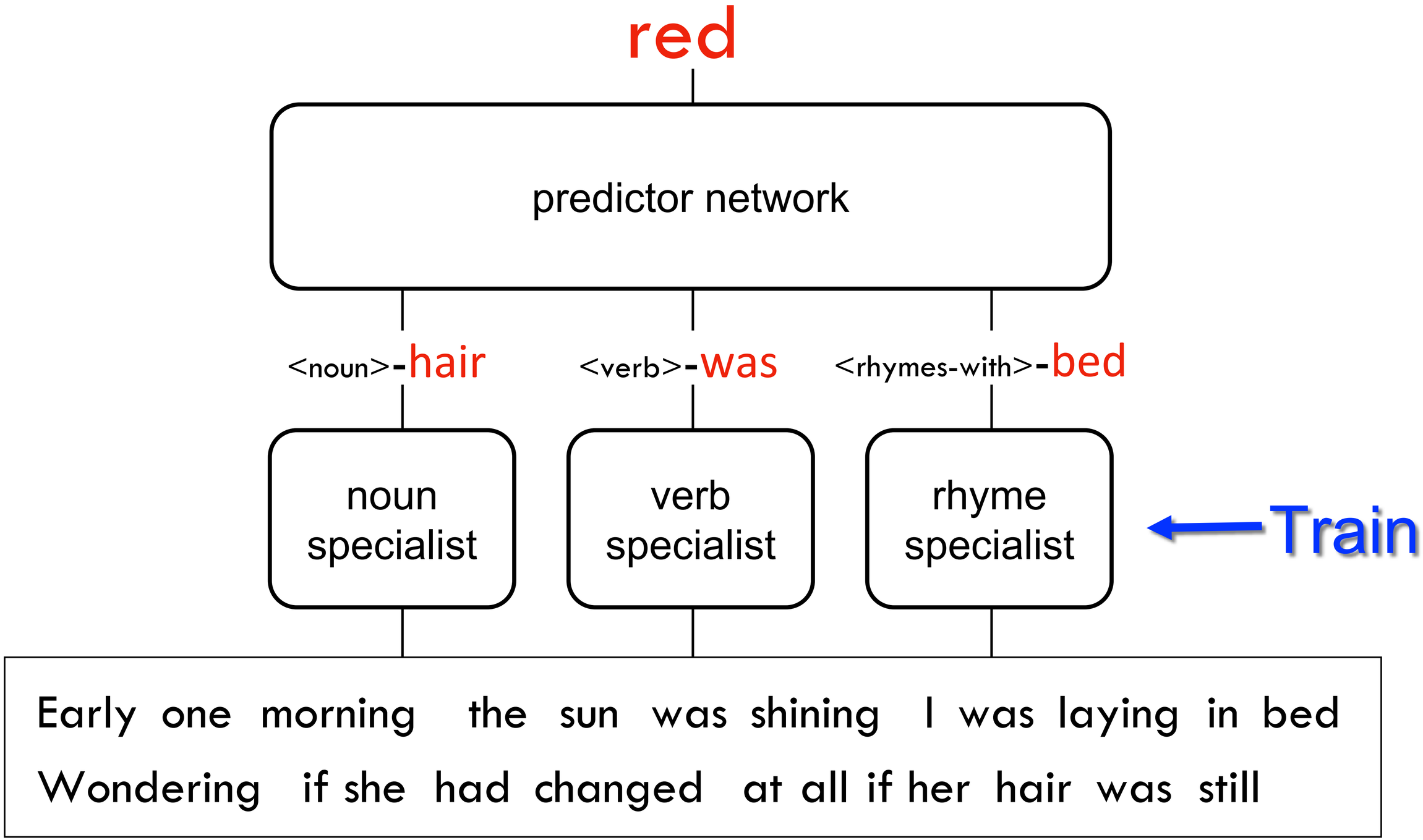

<rhymes-with>-bed
rhyme specialist
Early one morning the sun was shining I was laying in bed
Wondering if she had changed at all if her hair was still

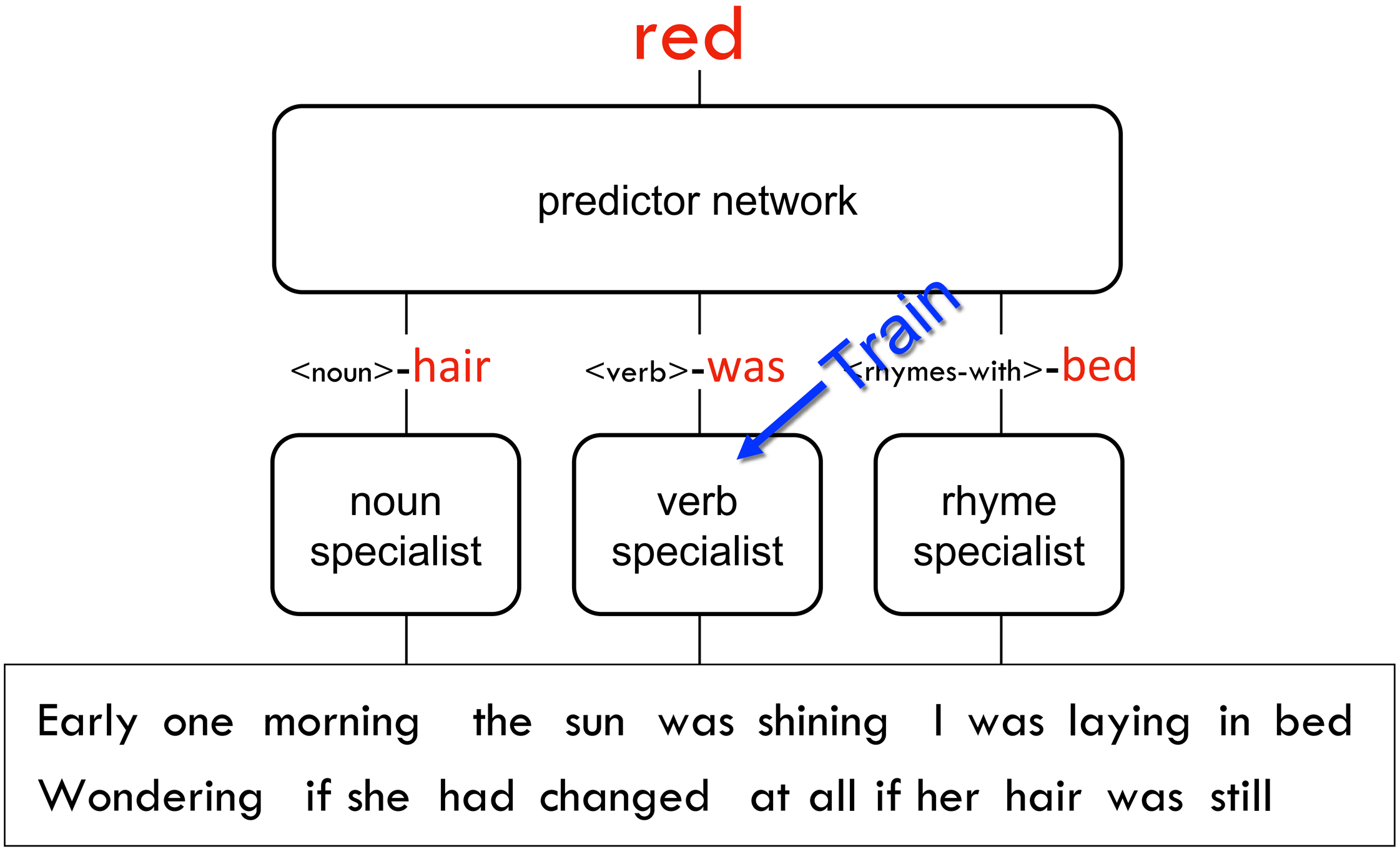

<noun>-hair
noun specialist
Early one morning the sun was shining I was laying in bed
Wondering if she had changed at all if her hair was still

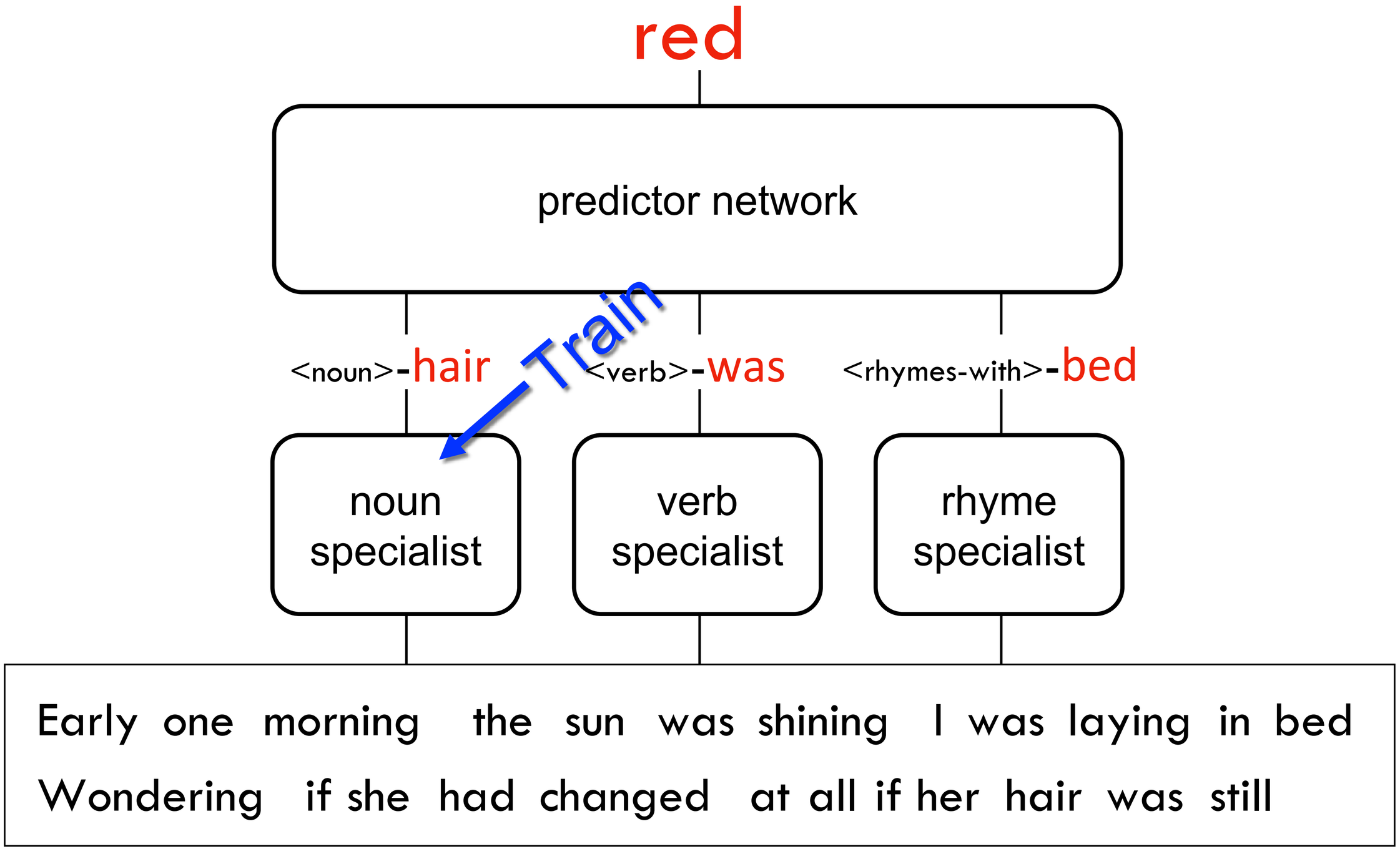

<verb>-was
verb specialist
Early one morning the sun was shining I was laying in bed
Wondering if she had changed at all if her hair was still

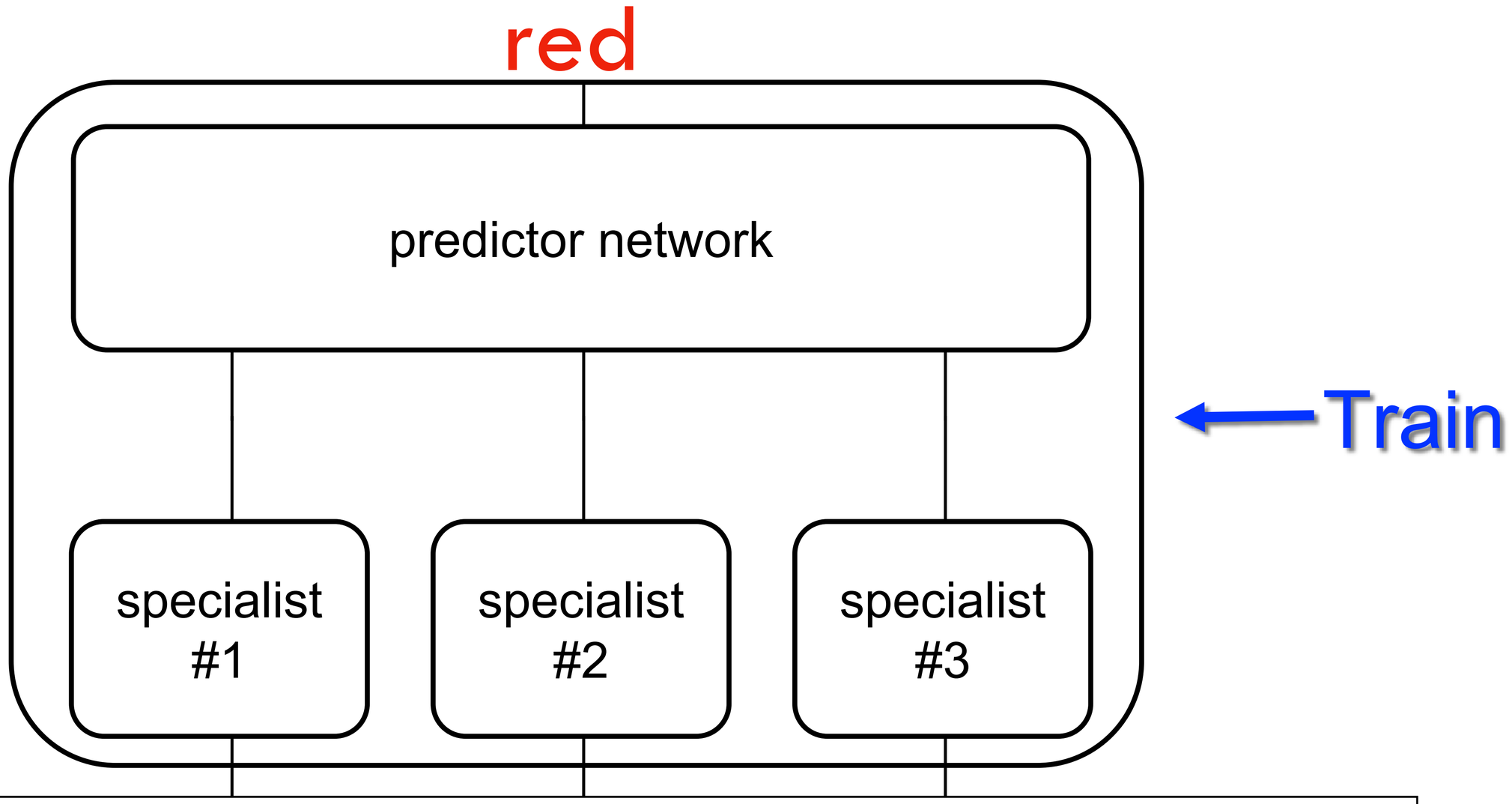

red
predictor network
Train
<noun>-hair
<verb>-was
<rhymes-with>-bed
noun specialist
verb specialist
rhyme specialist
Early one morning the sun was shining I was laying in bed
Wondering if she had changed at all if her hair was still


red
predictor network
<noun>-hair
<verb>-was
<rhymes-with>-bed
noun specialist
verb specialist
rhyme specialist
Train
Early one morning the sun was shining I was laying in bed
Wondering if she had changed at all if her hair was still


red
predictor network
<noun>-hair
<verb>-was
<rhymes-with>-bed
noun specialist
verb specialist
rhyme specialist
Train
Early one morning the sun was shining I was laying in bed
Wondering if she had changed at all if her hair was still

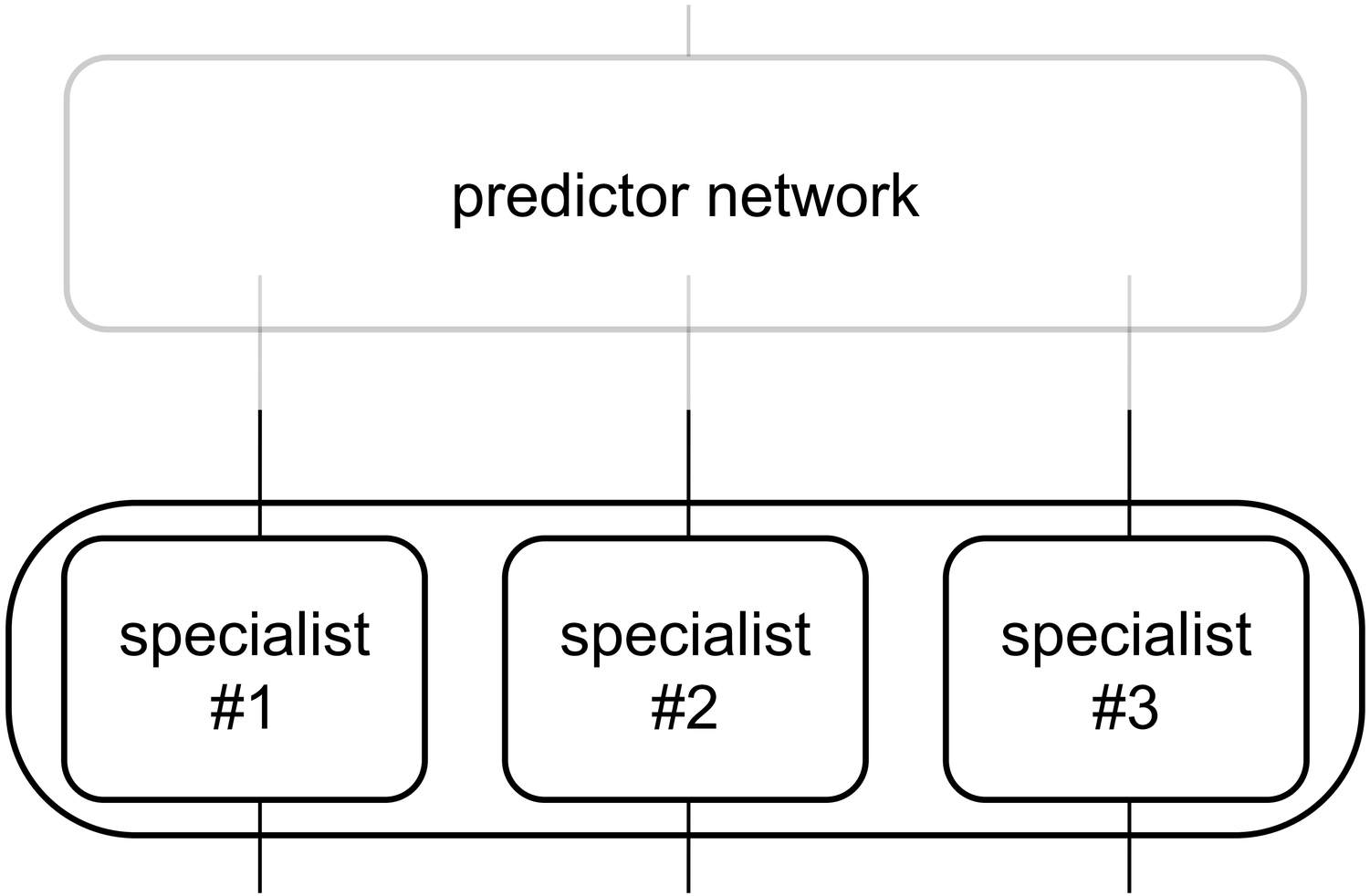

red
predictor network
<noun>-hair
<verb>-was
<rhymes-with>-bed
Train
noun specialist
verb specialist
rhyme specialist
Early one morning the sun was shining I was laying in bed
Wondering if she had changed at all if her hair was still


red
predictor network
Train
<noun>-hair
<verb>-was
<rhymes-with>-bed
noun specialist
verb specialist
rhyme specialist
Early one morning the sun was shining I was laying in bed
Wondering if she had changed at all if her hair was still

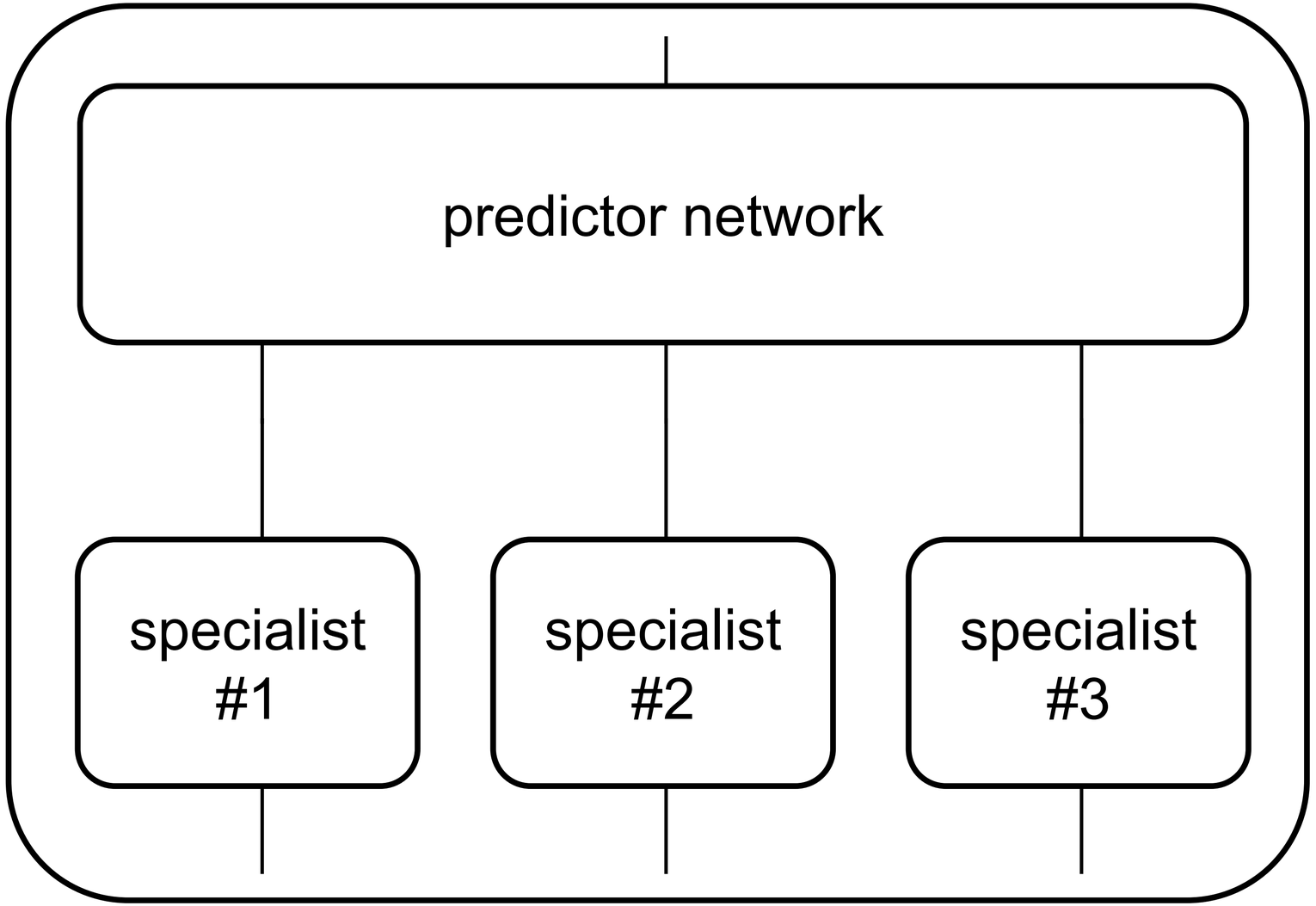



Early one morning the sun was shining I was laying in bed

Wondering if she had changed at all if her hair was still




# Attention Layer




# Transformer

The 16th President was ?

The capital of Zimbabwe is ?

Frank Zappa's middle name is ?

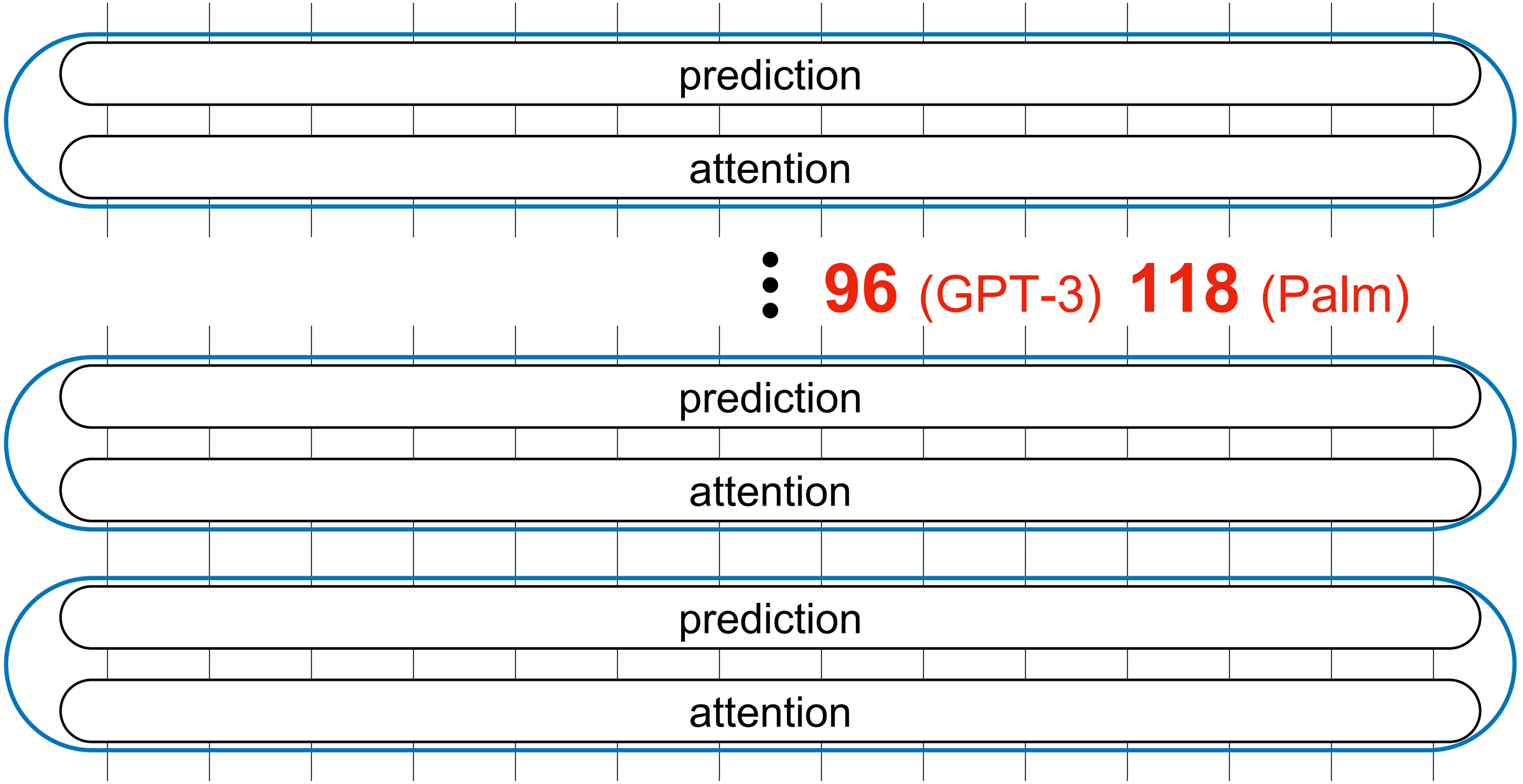
Napoleon was born on this date ?

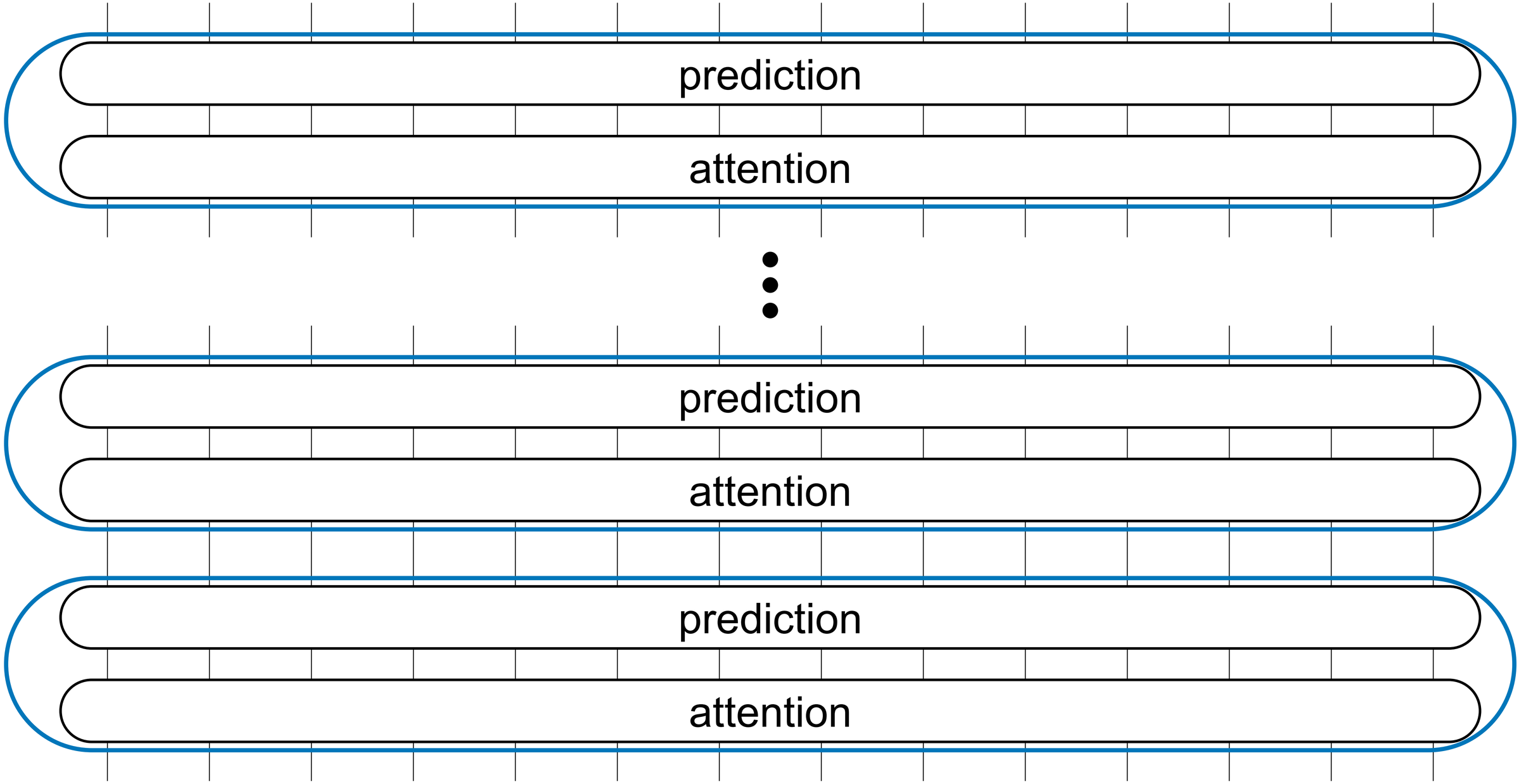
The prime factorization of 19456721434 is ?

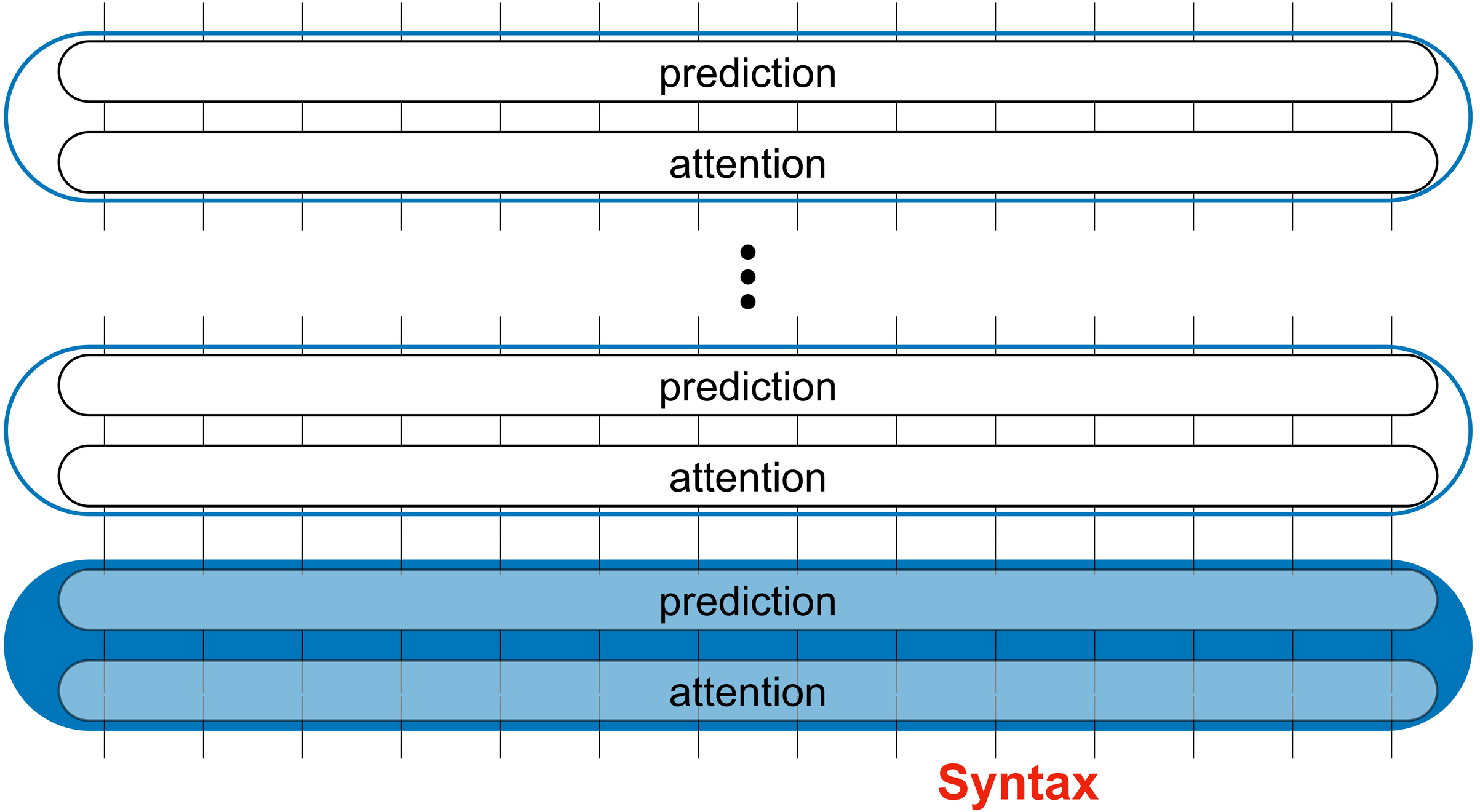
Queen Victoria's maiden name was ?

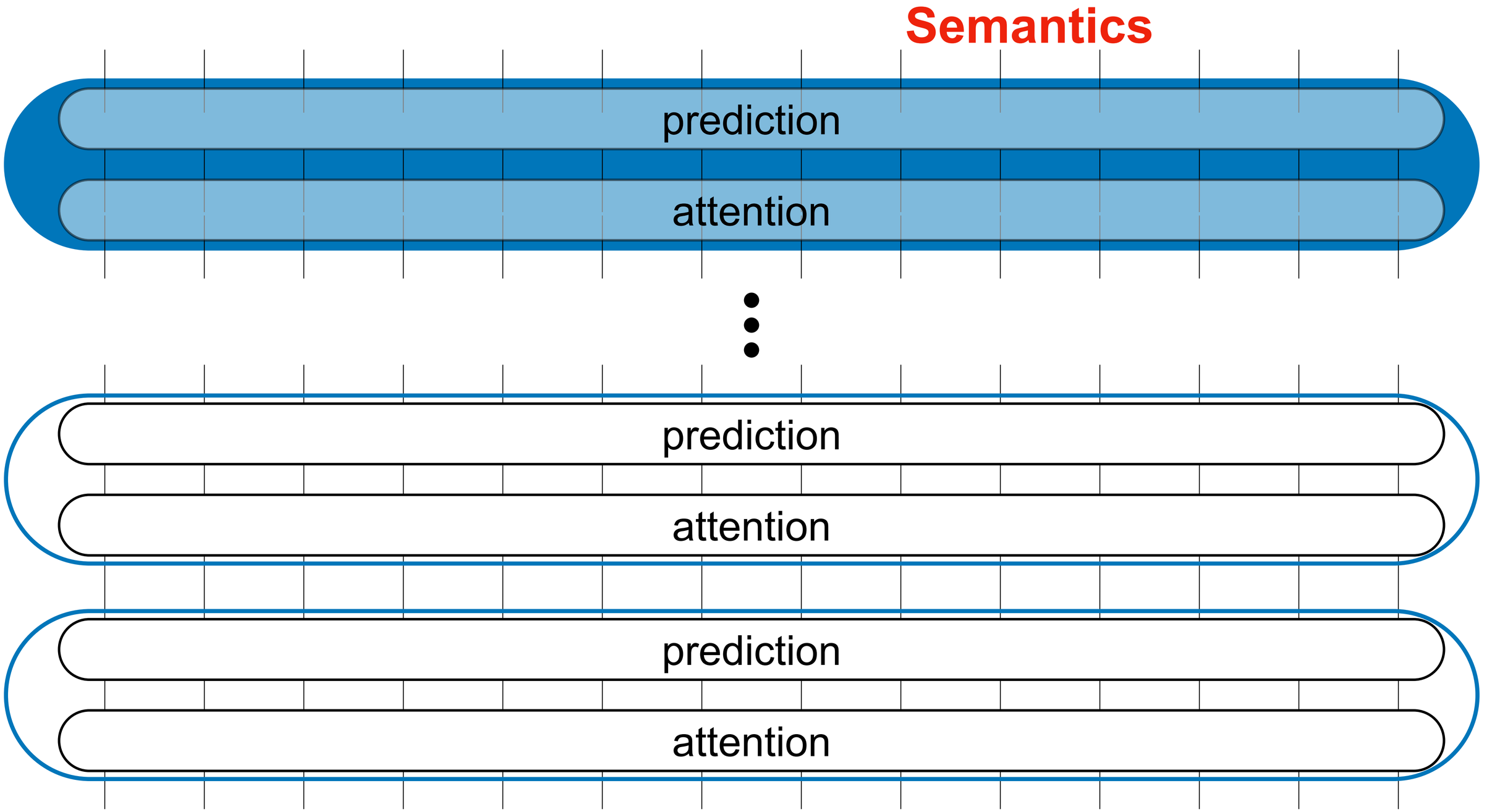
US per-capita income in 1957 was ?

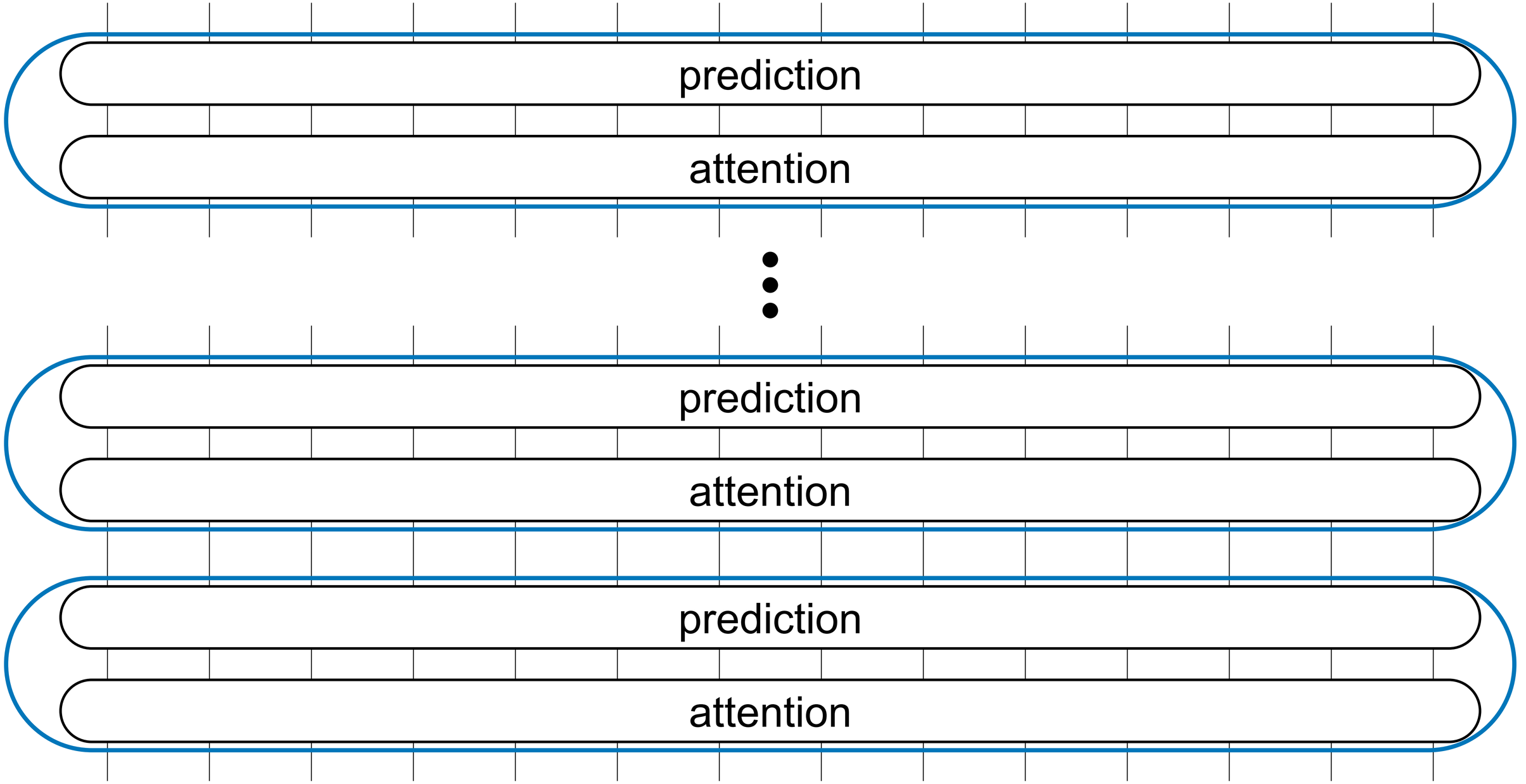
The lat long coordinates of Rome are ?

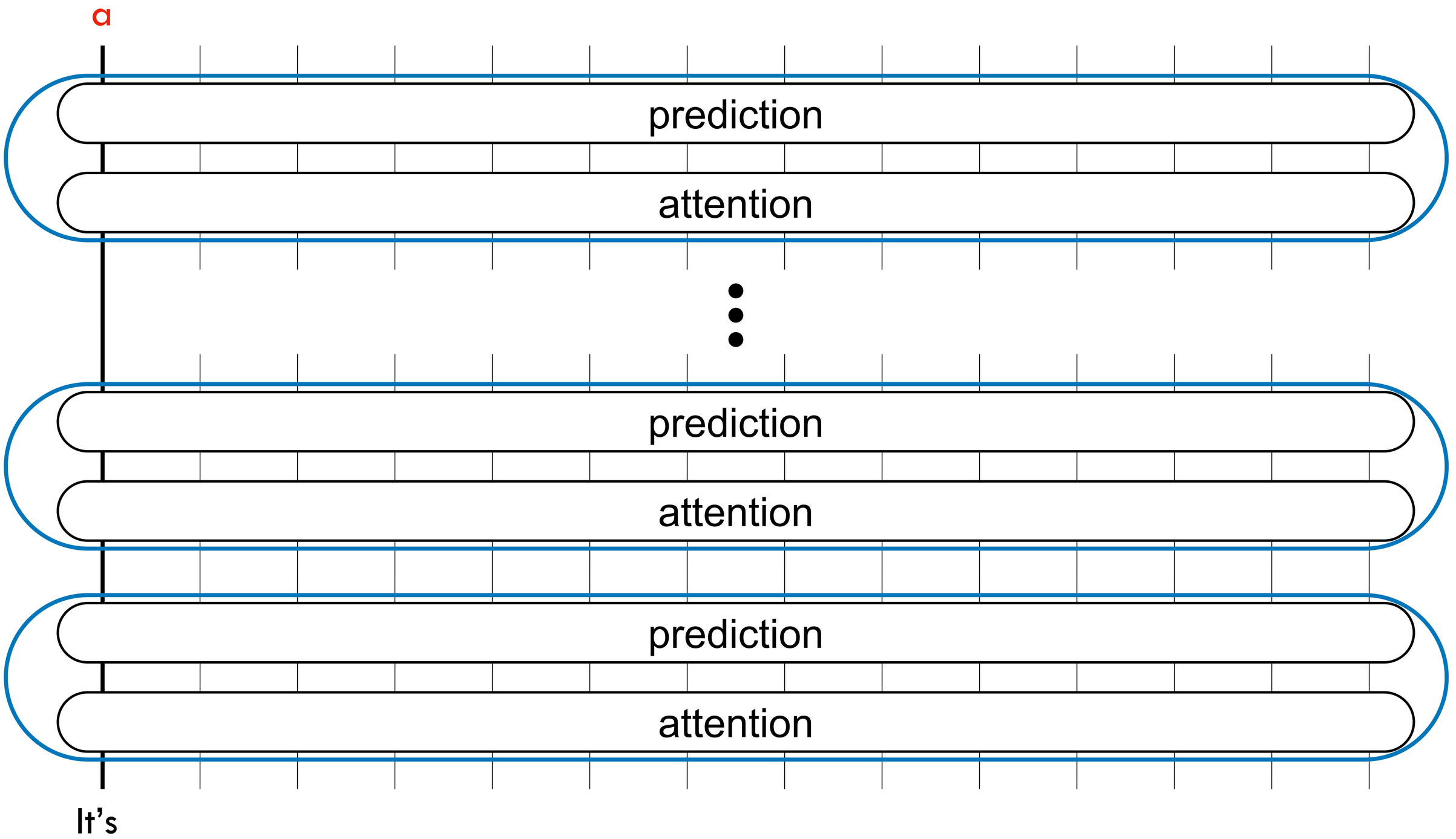

prediction
attention
96 (GPT-3) 118 (Palm)
prediction
attention
prediction
attention

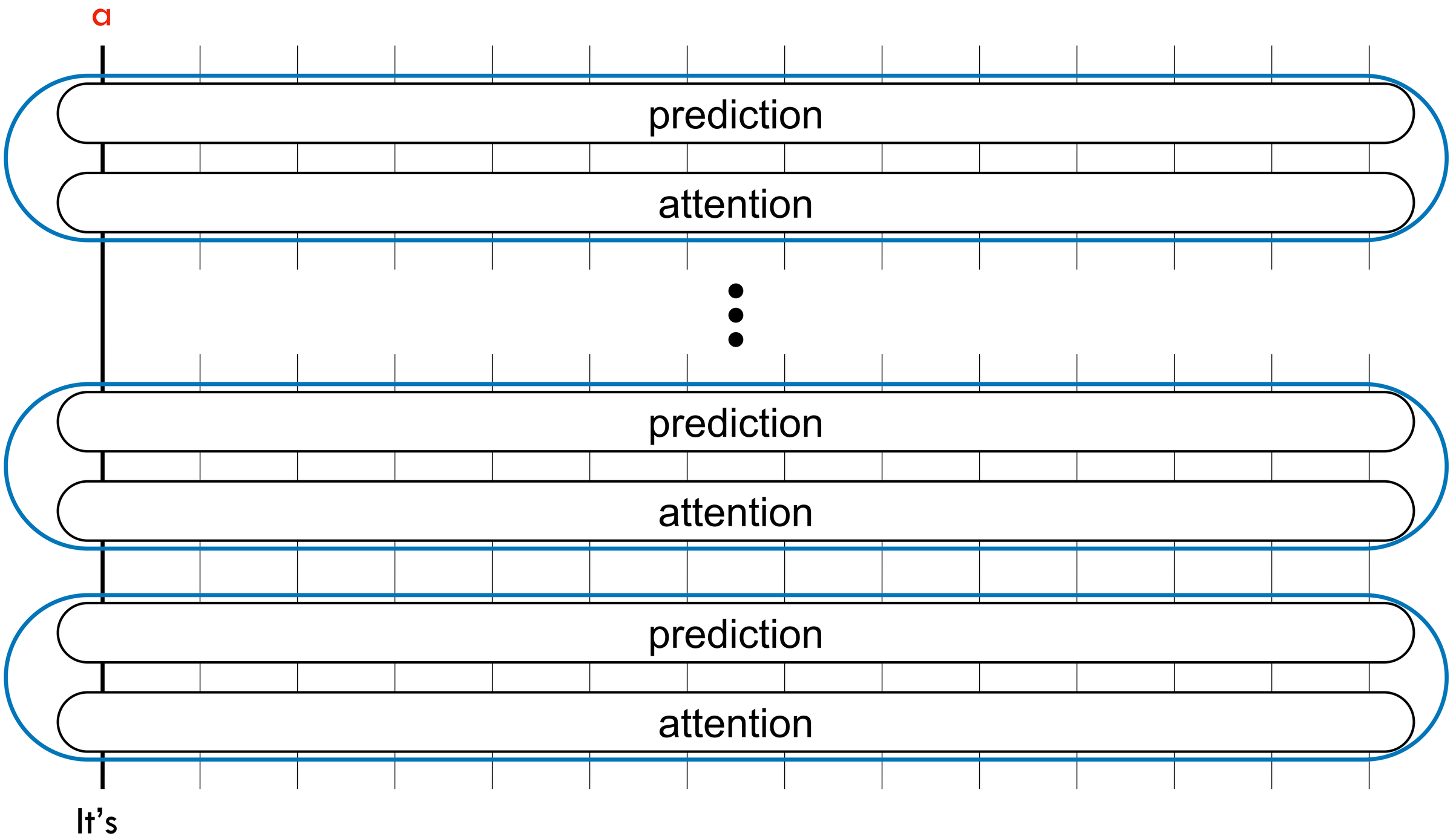

prediction
attention
prediction
attention
prediction
attention

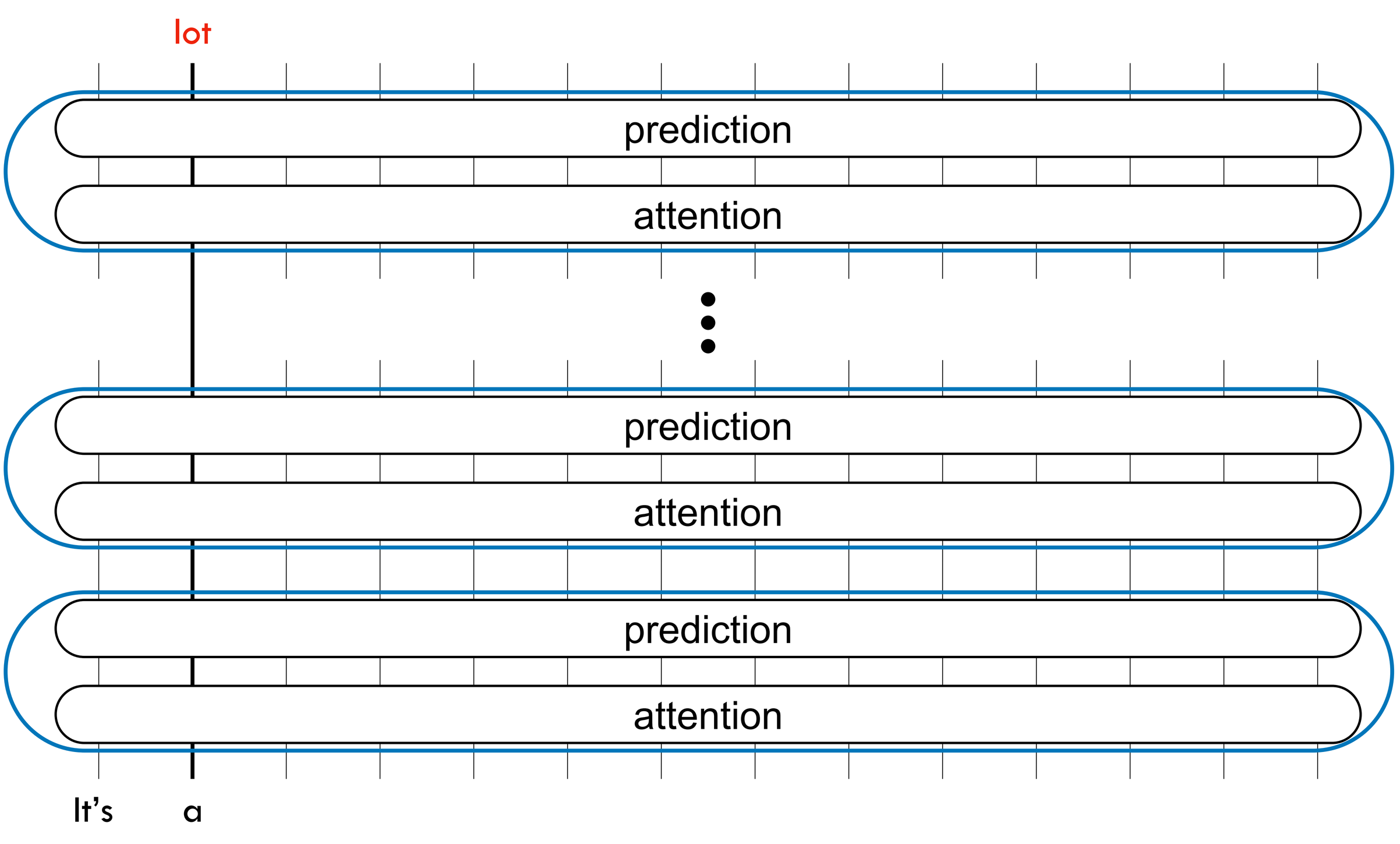

prediction
attention
prediction
attention
prediction
attention
Syntax

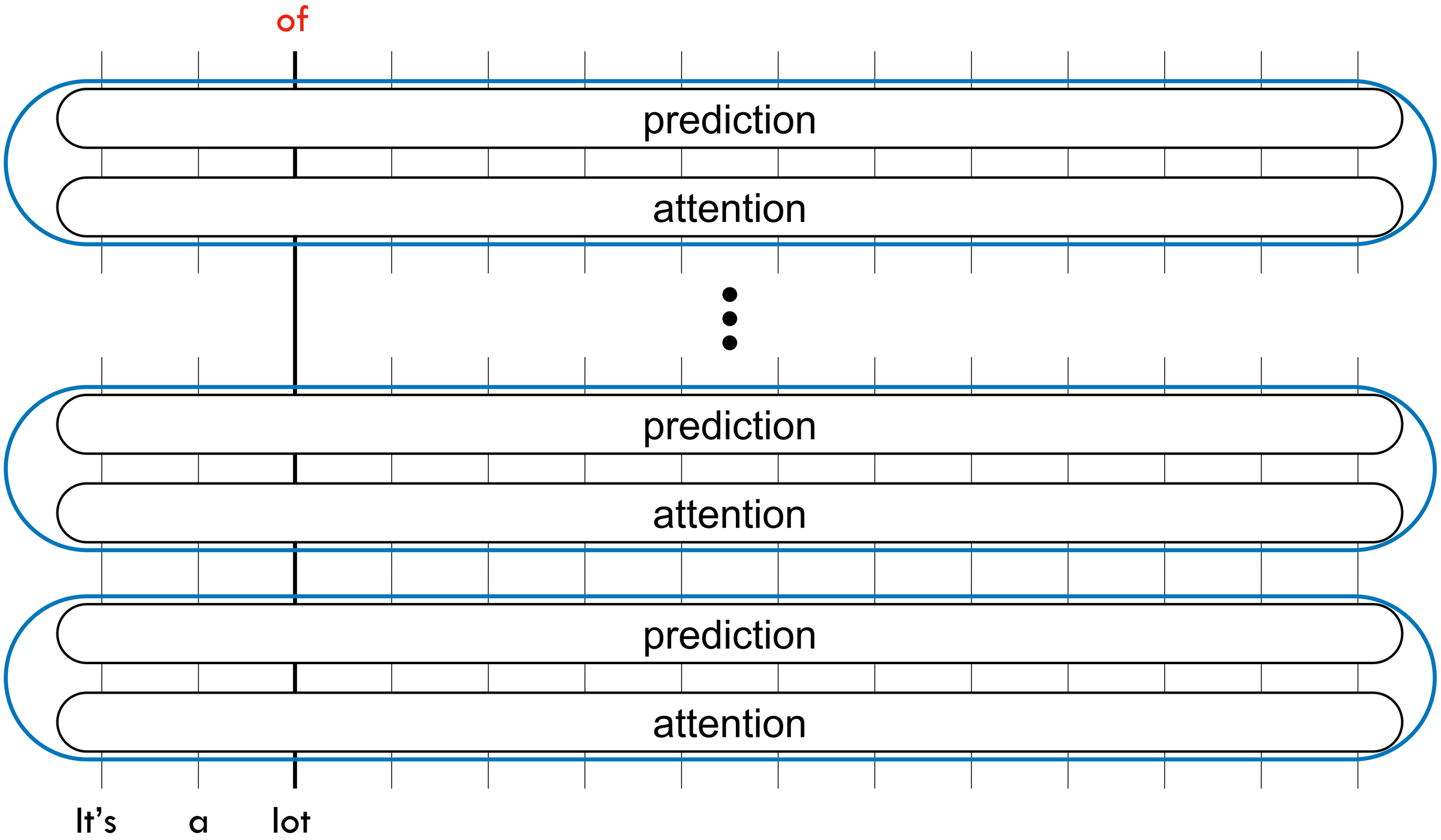

Semantics
prediction
attention
prediction
attention
prediction
attention

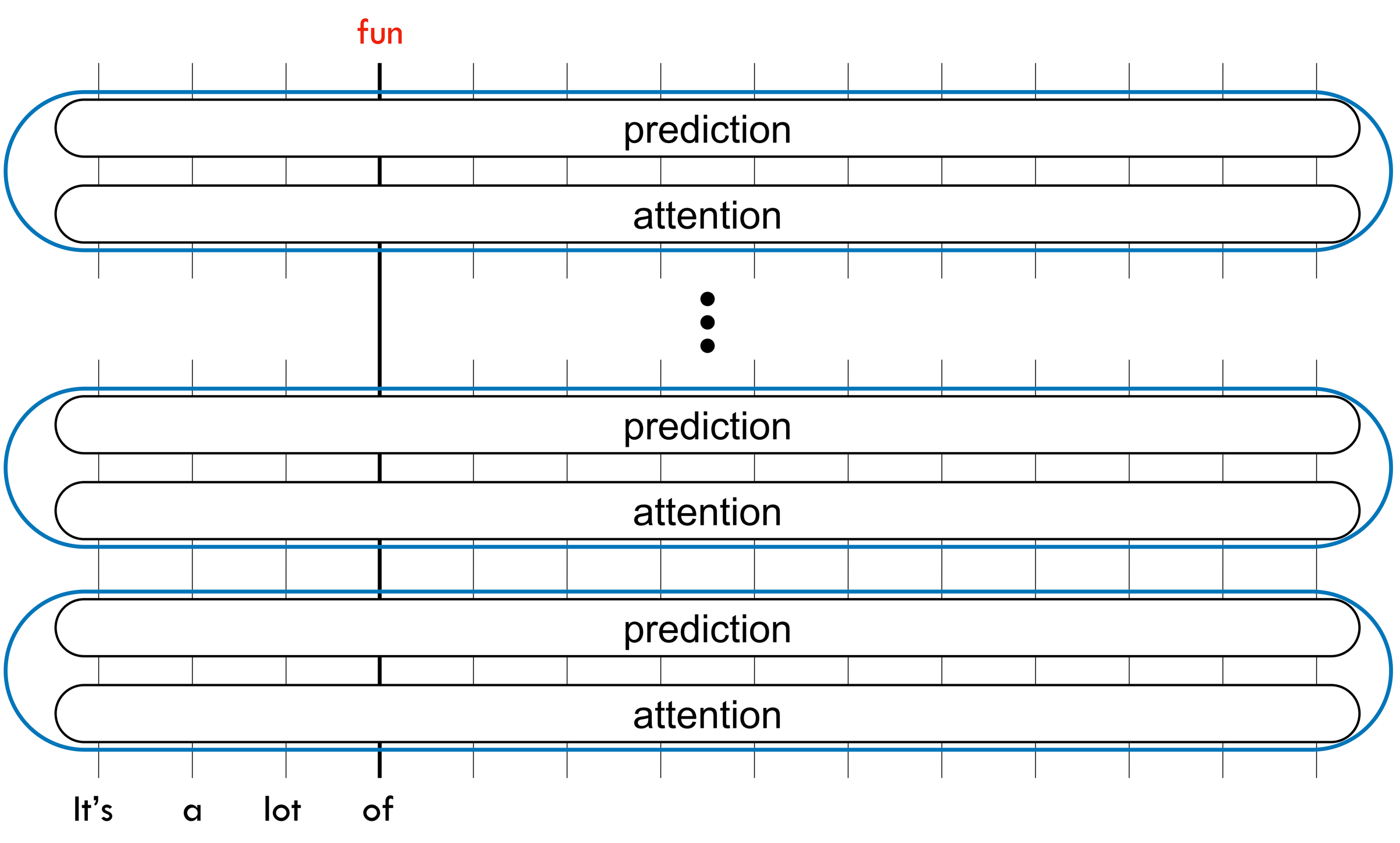

prediction
attention
prediction
attention
prediction
attention

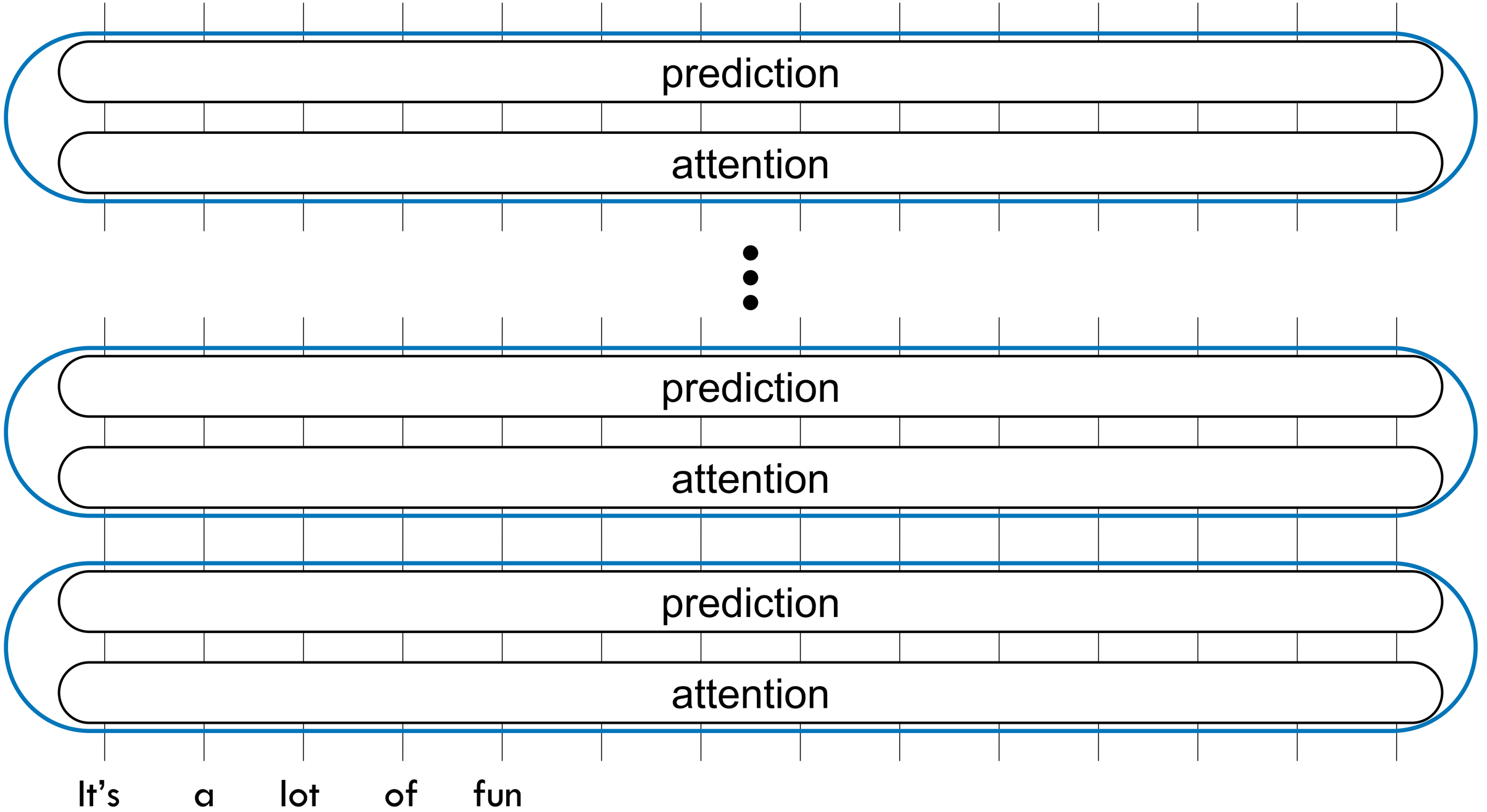

a
prediction
attention
prediction
attention
prediction
attention
It's

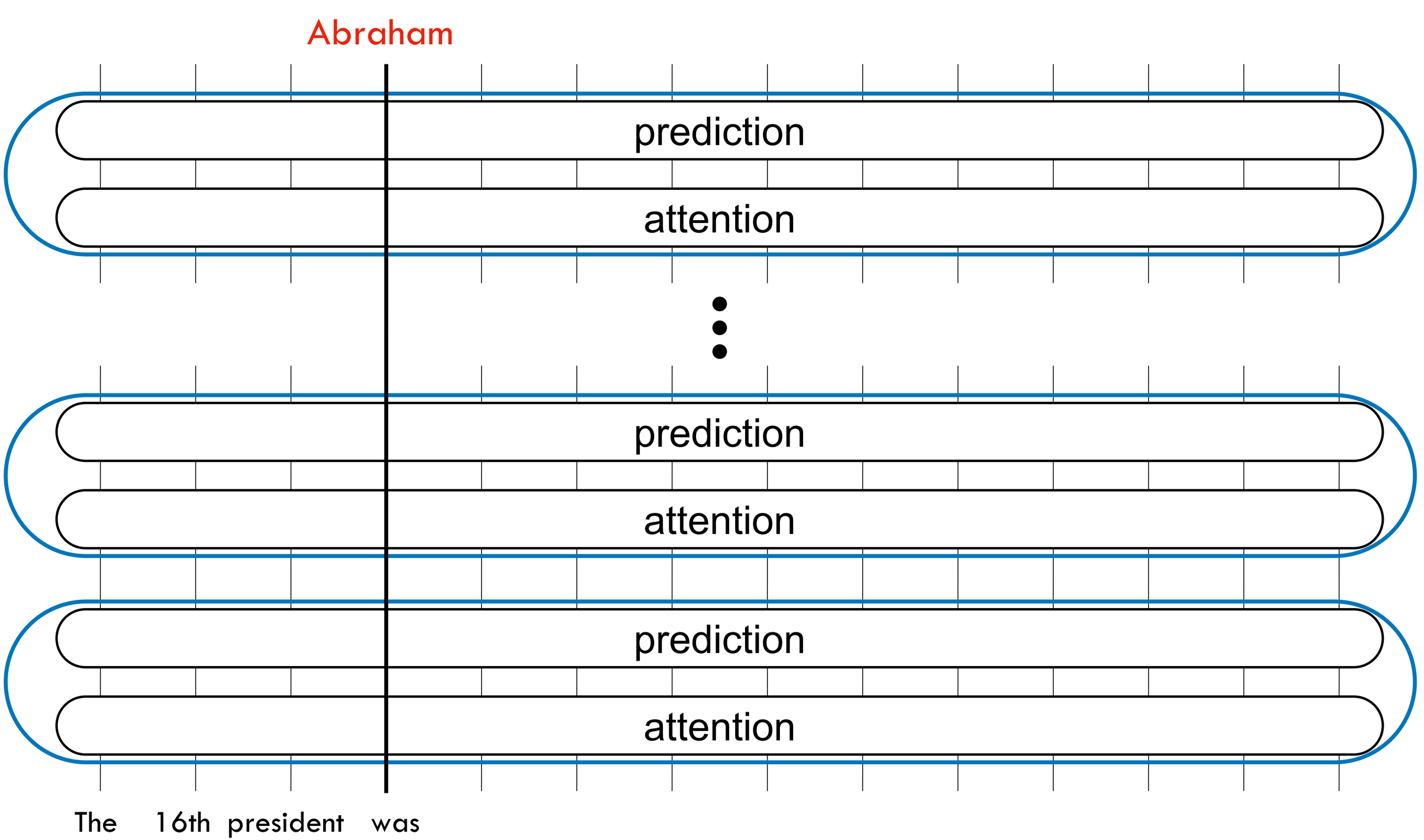

a
prediction
attention
prediction
attention
prediction
attention
It's

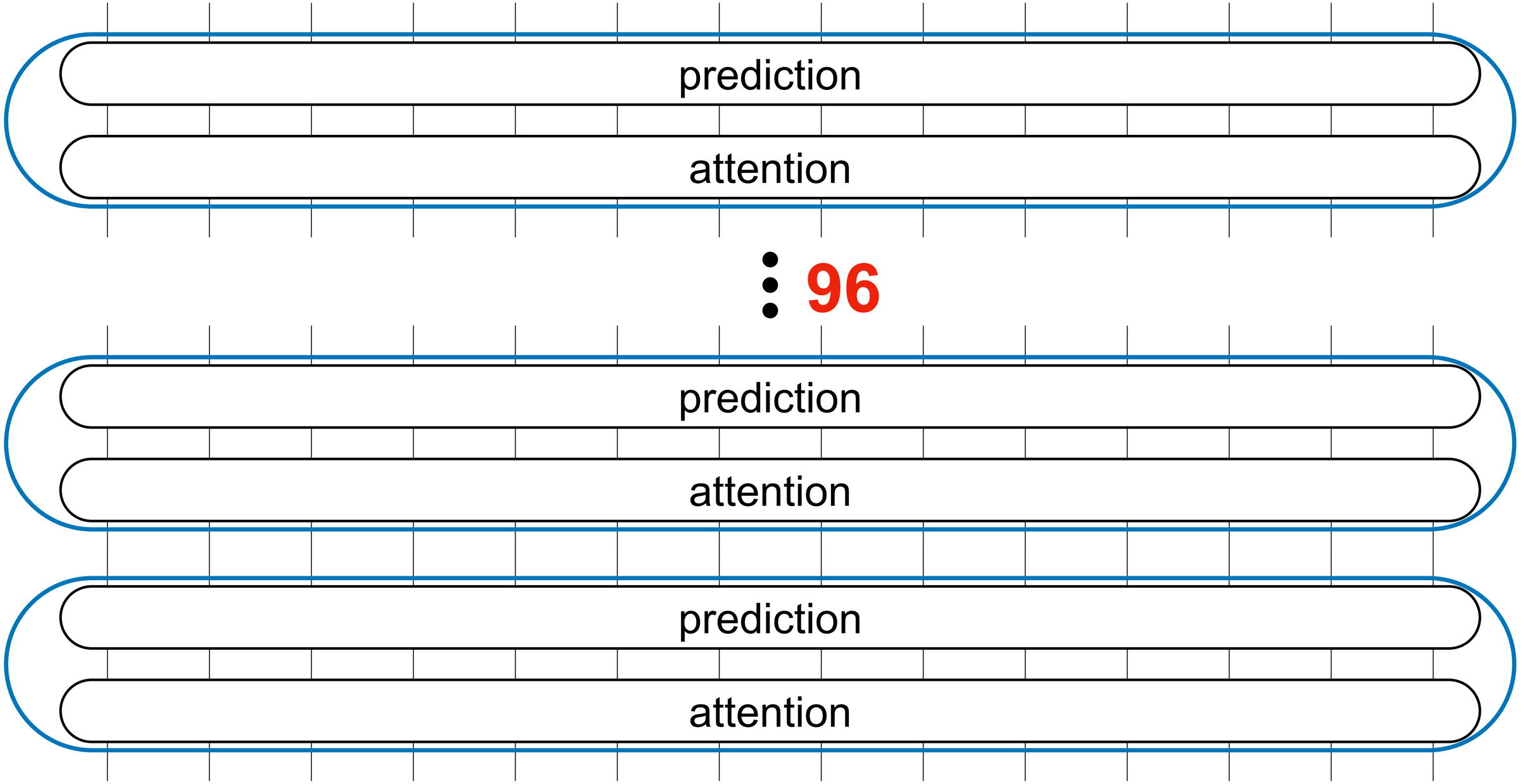

lot
prediction
attention
prediction
attention
prediction
attention
It's
a

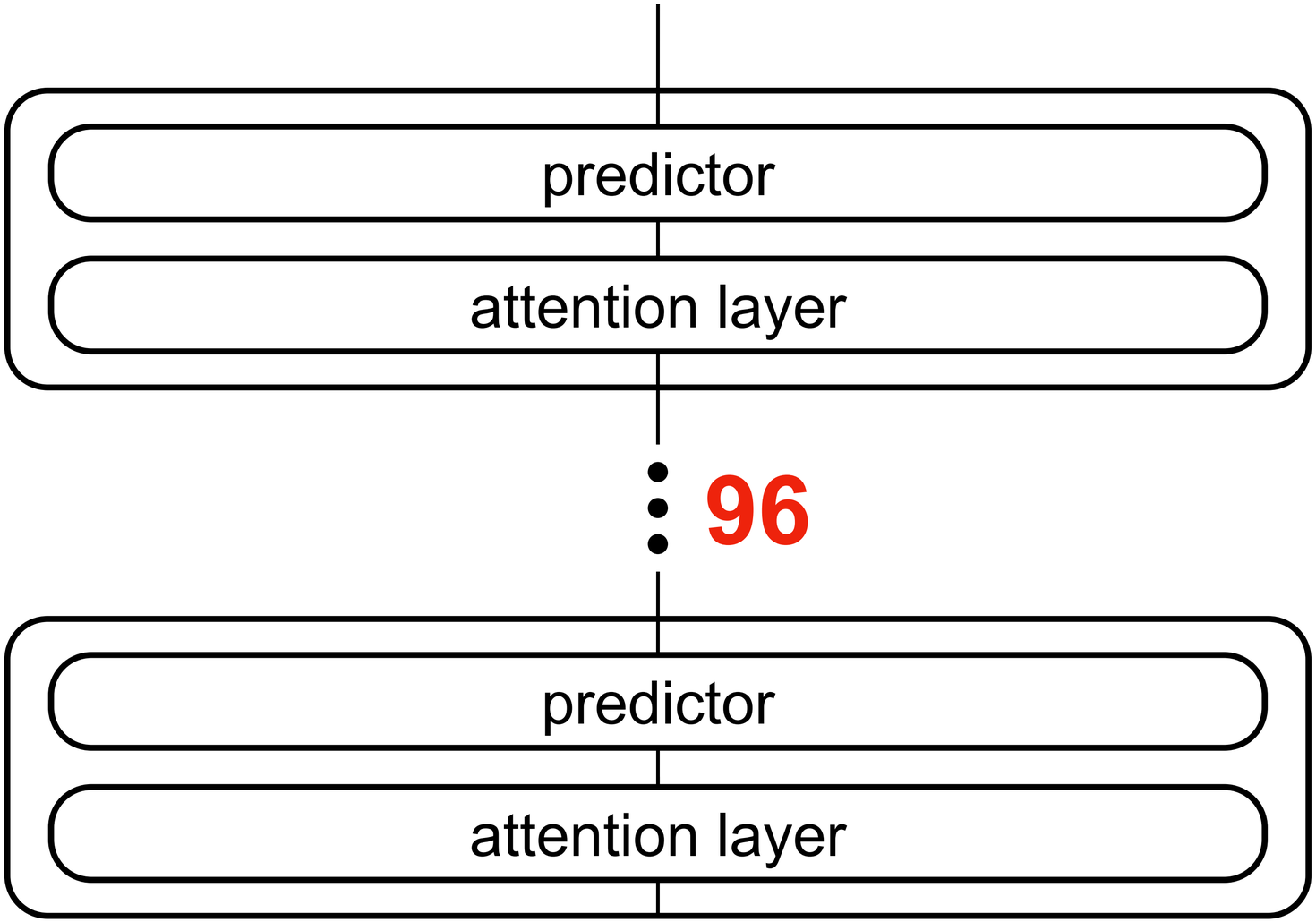

of
prediction
attention
prediction
attention
prediction
attention
It's
a
lot


fun
prediction
attention
prediction
attention
prediction
attention
It's
a
lot
of


prediction
attention
prediction
attention
prediction
attention
It's
a
lot
of
fun


Abraham
prediction
attention
prediction
attention
prediction
attention
The 16th president was


prediction
attention
96
prediction
attention
prediction
attention




**input size: 2048 words**

# GPT-3

**1 GPU**

355 years

**1000s of GPUs**

a month

The 16th President was

The capital of Zimbabwe is

Frank Zappa's middle name is

Napoleon was born on this date

The prime factorization of 19456721434 is

Queen Victoria's maiden name was

US per-capita income in 1957 was

The lat long coordinates of Rome are

The 16th President was Abraham Lincoln

The capital of Zimbabwe is Harare

Frank Zappa's middle name is Vincent

Napoleon was born on this date 1769

The prime factorization of 19456721434 is 2×3×3×17

Queen Victoria's maiden name was Alexandrina Victoria

US per-capita income in 1957 was $2,974

The lat long coordinates of Rome are 41.894722, 12.48

# Towards Understanding and Mitigating Social Biases in Language Models

**Paul Pu Liang** [1] **Chiyu Wu** [1] **Louis-Philippe Morency** [1] **Ruslan Salakhutdinov** [1]


## Abstract

*Warning: this paper contains model outputs that may be offensive or upsetting.*

As machine learning methods are deployed in real-world settings such as healthcare, legal systems, and social science, it is crucial to recognize how they shape social biases and stereotypes in these sensitive decision-making processes. Among such real-world deployments are large-scale pretrained language models (LMs) that can be potentially dangerous in manifesting undesirable *representational biases* - harmful biases resulting from stereotyping that propagate negative generalizations involving gender, race, religion, and other social constructs. As a step towards improving the fairness of LMs, we carefully define several sources of representational biases before proposing new benchmarks and metrics to measure them. With these tools, we propose steps towards mitigating social biases during text generation. Our empirical results and human evaluation demonstrate effectiveness in mitigating bias while retaining crucial contextual information for high-fidelity text generation, thereby pushing forward the performance-fairness Pareto frontier.


## 1. Introduction

Machine learning tools for processing large datasets are increasingly deployed in real-world scenarios such as healthcare (Velupillai et al., 2018), legal systems (Dale, 2019), and computational social science (Bamman et al., 2016). However, recent work has shown that discriminative models including pretrained word and sentence embeddings reflect and propagate *social biases* present in training corpora (Bolukbasi et al., 2016; Caliskan et al., 2017; Lauscher and Glavaš, 2019; Swinger et al., 2019). Further usages of such approaches can amplify biases and unfairly discriminate against users, particularly those from disadvantaged social groups (Barocas and Selbst, 2016; Sun et al., 2019; Zhao et al., 2017). More recently, language models (LMs) are increasingly used in real-world applications such as text generation (Radford et al., 2019), dialog systems (Zhang et al., 2020), recommendation systems (Shakespeare et al., 2020), and search engines (Baeza-Yates, 2016; Otterbacher et al., 2018). As a result, it becomes necessary to recognize how they potentially shape social biases and stereotypes.

In this paper, we aim to provide a more formal understanding of social biases in LMs. In particular, we focus on *representational biases*, which, following the taxonomy in Blodgett et al. (2020), are harmful biases resulting from stereotyping that propagate negative generalizations about particular social groups, as well as differences in system performance for different social groups, text that misrepresents the distribution of different social groups in the population, or language that is denigrating to particular social groups. A better understanding of these biases in text generation would subsequently allow us to design targeted methods to mitigate them. We begin by summarizing three inherent difficulties in *defining* and *measuring* biases during text generation:

**P1 Granularity:** In prior work studying biases in embeddings, social biases are measured using a set of association tests between predefined social constructs (e.g., gender and racial terms) and social professions (e.g., occupations, academic fields). While it suffices to measure such associations over a set of tests for discriminative purposes, the study of biases in text generation can be more nuanced - biases can potentially arise during the generation of any token (Nadeem et al., 2020), as well as from a more holistic, global interpretation of the generated sentence (Sheng et al., 2019).

**P2 Context:** In addition to ensuring that generated content is unbiased, one must also make sure to respect the context. Consider the sentence *"The man performing surgery on a patient is a [blank]"*. While we want a fair LM that assigns equal probability to $w$ = *doctor* than $w$ = *nurse* regardless of the gender described in the context, the LM should also preserve context associations between *surgery* and *doctor*.

**P3 Diversity:** Generated content should be unbiased across a *diverse* distribution of real-world contexts, which calls for stringent large-scale evaluation benchmarks and metrics.

Our first contribution is therefore to *disentangle* two sources

[1]Carnegie Mellon University. Correspondence to: Paul Pu Liang <pliang@cs.cmu.edu>.

# StereoSet: Measuring stereotypical bias in pretrained language models

**Moin Nadeem**[§] and **Anna Bethke**[†] and **Siva Reddy**[‡]
[§]Massachusetts Institute of Technology, Cambridge MA, USA
[†]Facebook, Menlo Park CA, USA
[‡]Facebook CIFAR AI Chair, Mila; McGill University, Montreal, QC, Canada
mnadeem@mit.edu anna.bethke@intel.com,
siva.reddy@mila.quebec


## Abstract

A stereotype is an over-generalized belief about a particular group of people, e.g., *Asians are good at math* or *African Americans are athletic*. Such beliefs (biases) are known to hurt target groups. Since pretrained language models are trained on large real world data, they are known to capture stereotypical biases. It is important to quantify to what extent these biases are present in them. Although this is a rapidly growing area of research, existing literature lacks in two important aspects: 1) they mainly evaluate bias of pretrained language models on a small set of artificial sentences, even though these models are trained on natural data; 2) current evaluations focus on measuring bias without considering the language modeling ability of a model, which could lead to misleading trust on a model even if it is a poor language model. We address both these problems. We present StereoSet, a large-scale natural English dataset to measure stereotypical biases in four domains: gender, profession, race, and religion. We contrast both stereotypical bias and language modeling ability of popular models like BERT, GPT2, RoBERTa, and XLNet. We show that these models exhibit strong stereotypical biases. Our data and code are available at https://stereoset.mit.edu.


**Choose the appropriate word:**

**Domain:** Gender **Target:** Girl

**Context:** Girls tend to be more ____ than boys

**Option 1:** soft (stereotype)
**Option 2:** determined (anti-stereotype)
**Option 3:** fish (meaningless)

(a) The Intrasentence Context Association Test

**Choose the appropriate sentence:**

**Domain:** Race **Target:** Arab

**Context:** He is an Arab from the Middle East.
**Option 1:** He is probably a terrorist with bombs. (stereotype)
**Option 2:** He is a pacifist. (anti-stereotype)
**Option 3:** My dog wants a walk. (meaningless)

(b) The Intersentence Context Association Test

Figure 1: Context Association Tests (CATs) measure both bias and language modeling ability of language models.

## 1 Introduction

A key idea behind the current success of neural network models for language is pretrained representations such as word embeddings (Mikolov

pora using variants of language modeling objective (i.e., predicting a word given its surrounding context). In the recent years, these representations empowered neural models to attain unprecedented levels of performance gains on multiple language tasks. The resulting models are being deployed widely as services on platforms like Google Cloud and Amazon AWS to serve millions of users.

While this growth is commendable, there are concerns about the fairness of these models. Since pretrained representations are obtained from learning on massive text corpora, there is a danger that stereotypical biases in the real world are reflected