Processing Language in Humans and Machines

Michael Hahn

May 17, 2024

A bit about myself

- Assistant Professor (W2TTW3) in Language Science & Technology

Image credit: uni-saarland.de; theory.stanford.edu; https://tuebingen.place/

A bit about myself

- Assistant Professor (W2TTW3) in Language Science & Technology

- 2022 PhD, Stanford University

Image credit: uni-saarland.de; theory.stanford.edu; https://tuebingen.place/

A bit about myself

- Assistant Professor (W2TTW3) in Language Science & Technology

- 2022 PhD, Stanford University
- Previously research stays at MIT (2019), Facebook AI Research (2018), Edinburgh (2016)

Image credit: uni-saarland.de; theory.stanford.edu; https://tuebingen.place/

A bit about myself

- Assistant Professor (W2TTW3) in Language Science & Technology

- 2022 PhD, Stanford University
- Previously research stays at MIT (2019), Facebook AI Research (2018), Edinburgh (2016)
- B.Sc. and M.Sc. in Mathematics in Tübingen (2015)

Image credit: uni-saarland.de; theory.stanford.edu; https://tuebingen.place/

Natural Language Processing

Natural Language Processing

1966: ELIZA

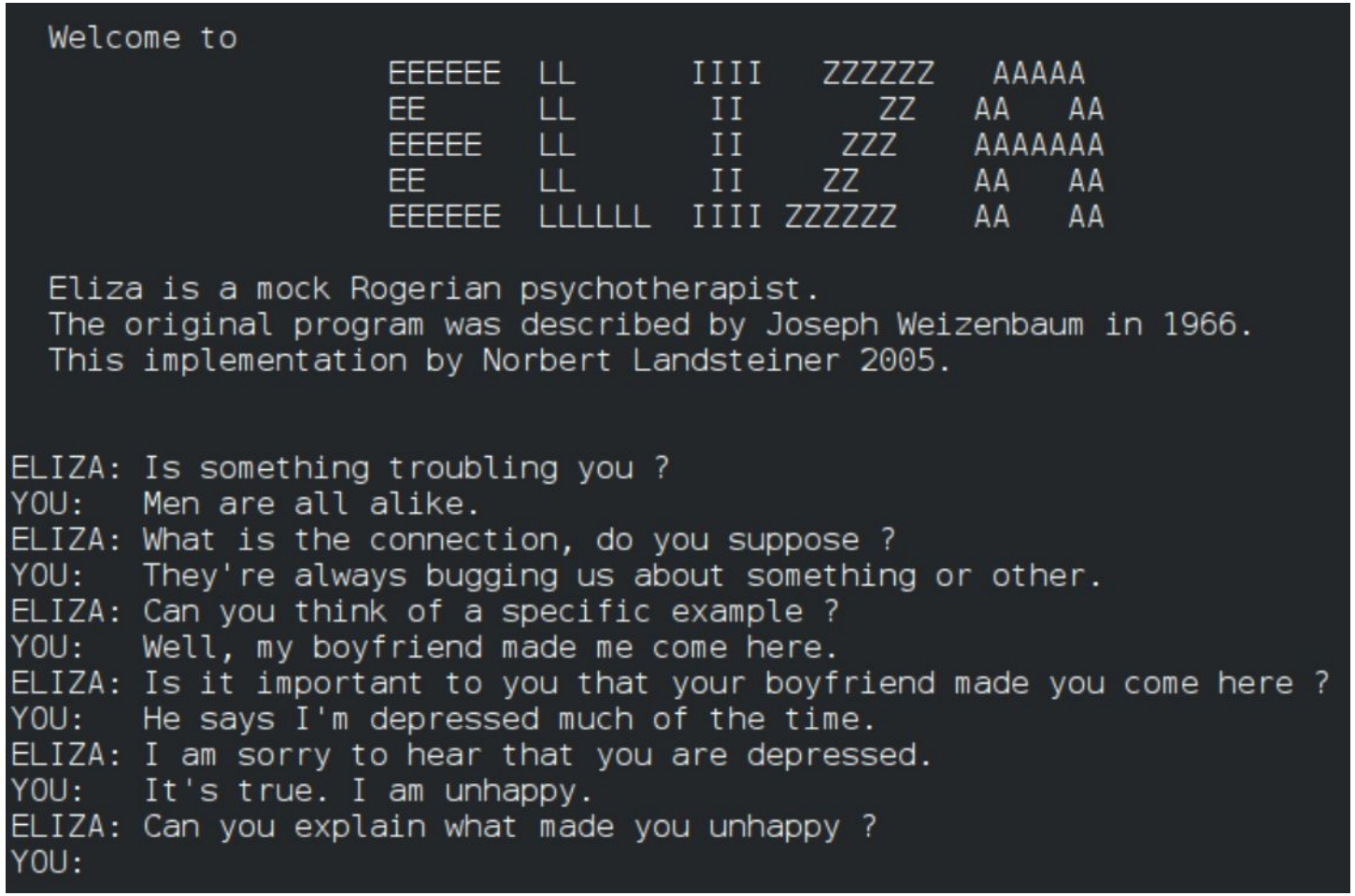


Image credit: https://sites.google.com/view/elizagen-org/the-original-eliza

Natural Language Processing

1966: ELIZA

```
Welcome to
                EEEEEE  LL      IIII   ZZZZZZ   AAAAA
                EE      LL       II       ZZ   AA   AA
                EEEEE   LL       II     ZZZ    AAAAAAA
                EE      LL       II    ZZ      AA   AA
                EEEEEE  LLLLLL  IIII ZZZZZZ    AA   AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
l alike.
e connection, do you suppose ?
ways bugging us about something or other.
ink of a specific example ?
oyfriend made me come here.
rtant to you that your boyfriend made you come here ?
m depressed much of the time.
 to hear that you are depressed.
 I am unhappy.
plain what made you unhappy ?
```

```
(DO YOU WISH TO BE ABLE TO 4) (DOES THAT TROUBLE YOU))                    000620
((0 YOU FEEL 0) (TELL ME MORE ABOUT SUCH FEELINGS)                        000630
(DO YOU OFTEN FEEL 4)                                                     000640
(DO YOU ENJOY FEELING 4)                                                  000650
(OF WHAT DOES FEELING 4 REMIND YOU))                                      000660
((0 YOU 0 I 0) (PERHAPS IN YOUR FANTASY WE 3 EACH OTHER)                  000670
(DO YOU WISH TO 3 ME)                                                     000680
(YOU SEEM TO NEED TO 3 ME)                                                000690
(DO YOU 3 ANYONE ELSE))                                                   000700
((0) (YOU SAY 1)                                                          000710
(CAN YOU ELABORATE ON THAT) (DO YOU SAY 1 FOR SOME SPECIAL REASON)        000720
(THAT'S QUITE INTERESTING)))                                              000730
(YOU = I ((0 I 0 YOU 0)                                                   000740
(WHY DO YOU THINK I 3 YOU)                                                000750
(DID YOUR PARENTS 3 YOU))                                                 000760
((0 I AM 0) (WHAT MAKES YOU THINK I AM 4)                                 000770
(DOES IT PLEASE YOU TO BELIEVE I AM 4)                                    000780
(DO YOU SOMETIMES WISH YOU WERE 4)                                        000790
(PERHAPS YOU WOULD LIKE TO BE 4))                                         000800
                                                                          000810
```

Image credit: https://sites.google.com/view/elizagen-org/the-original-eliza

Natural Language Processing

2011: Watson

Image credit: http://www.aaai.org/ojs/index.php/aimagazine/article/view/2303/2165

Natural Language Processing

2011: Watson

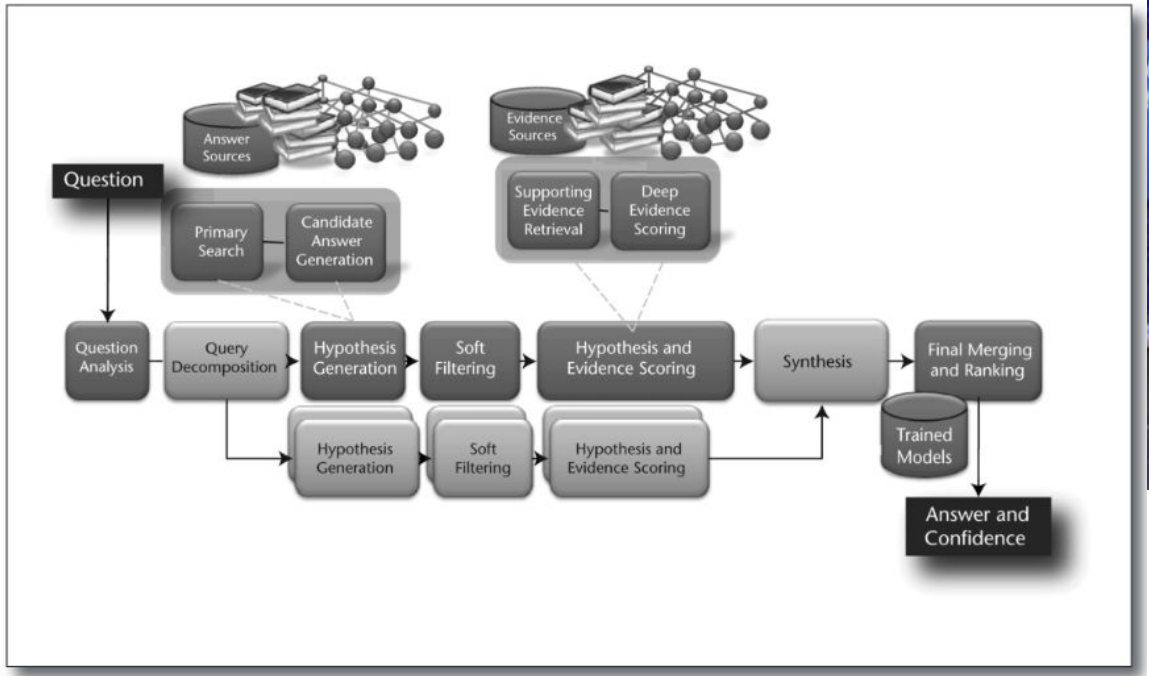
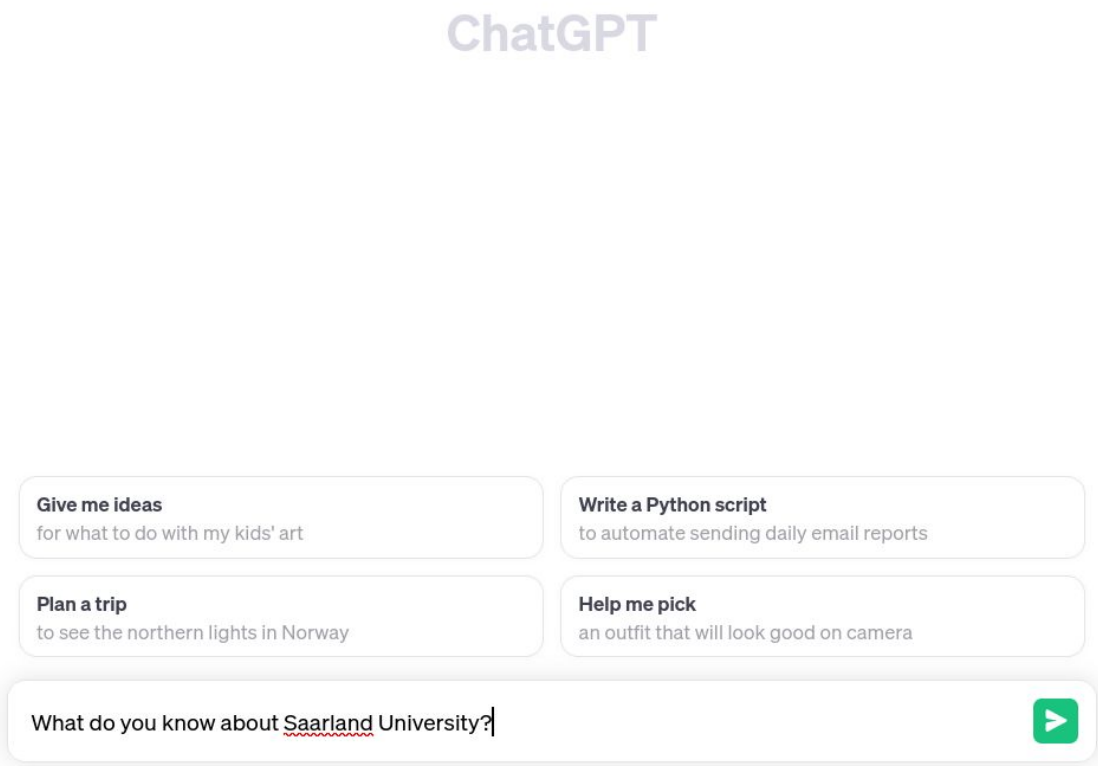


Image credit: http://www.aaai.org/ojs/index.php/aimagazine/article/view/2303/21

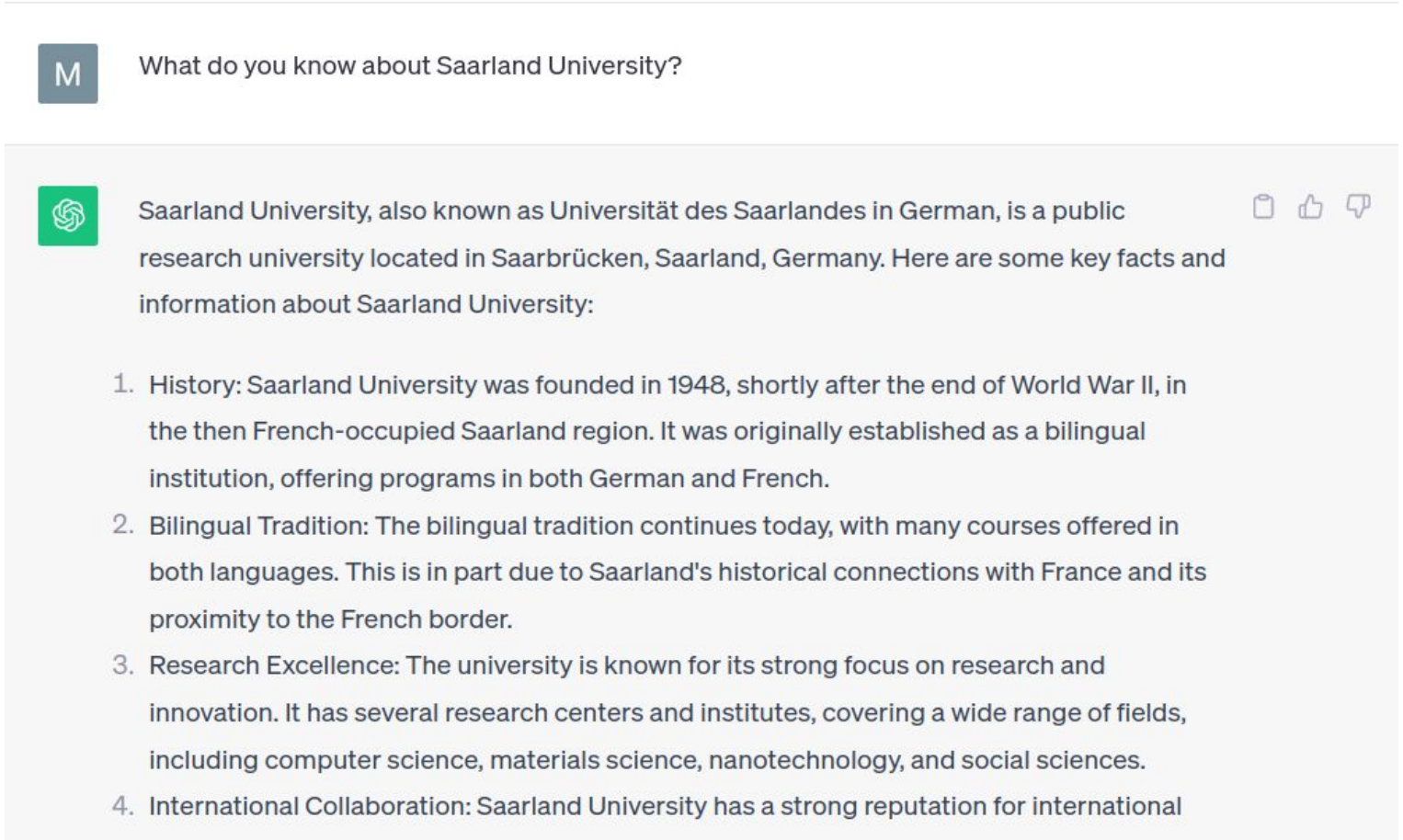
Natural Language Processing

2022: ChatGPT



Natural Language Processing

2022: ChatGPT



Natural Language Processing

2022: ChatGPT

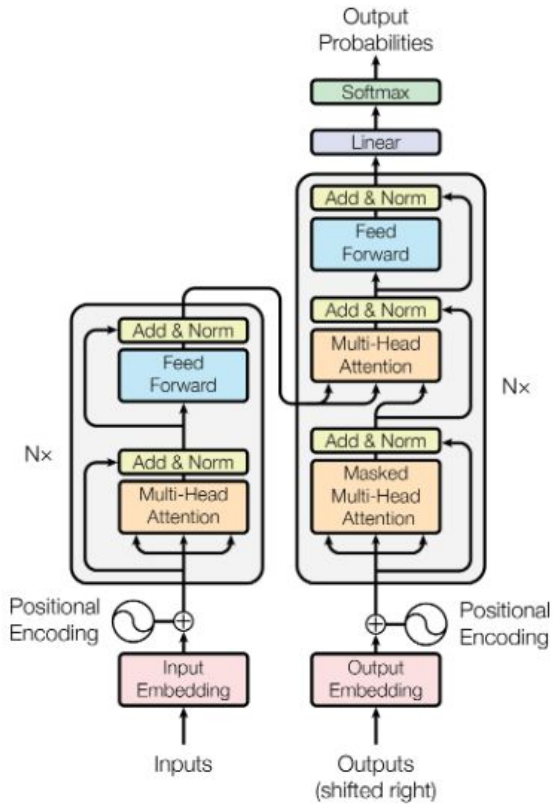


Image credit: Vaswani et al, https://arxiv.org/pdf/1706.03762.pdf

M

What do you know about Saarland University?

Saarland University, also known as Universität des Saarlandes in German, is a public research university located in Saarbrücken, Saarland, Germany. Here are some key facts and information about Saarland University:

1. History: Saarland University was founded in 1948, shortly after the end of World War II, in the then French-occupied Saarland region. It was originally established as a bilingual institution, offering programs in both German and French.
2. Bilingual Tradition: The bilingual tradition continues today, with many courses offered in both languages. This is in part due to Saarland's historical connections with France and its proximity to the French border.
3. Research Excellence: The university is known for its strong focus on research and innovation. It has several research centers and institutes, covering a wide range of fields, including computer science, materials science, nanotechnology, and social sciences.
4. International Collaboration: Saarland University has a strong reputation for international

Language Models

Autocomplete

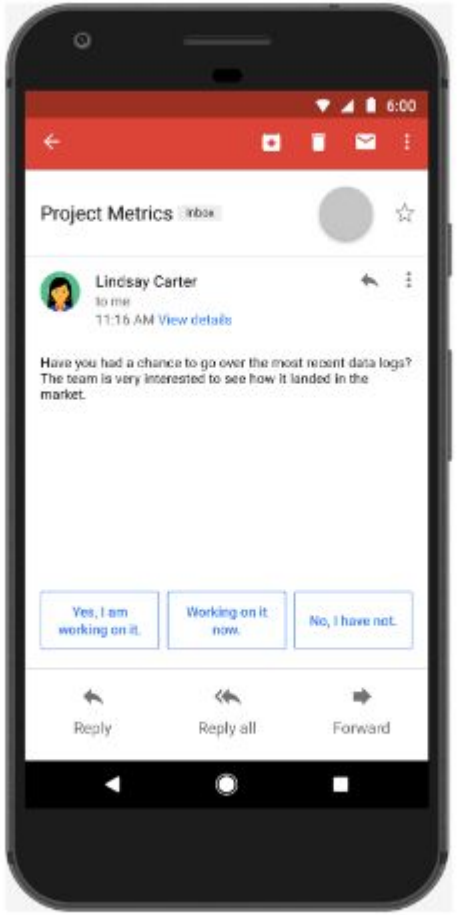


Image credit: https://blog.google/products/gmail/save-time-with-smart-reply-in-gmail/

Language Models

Helping people write code

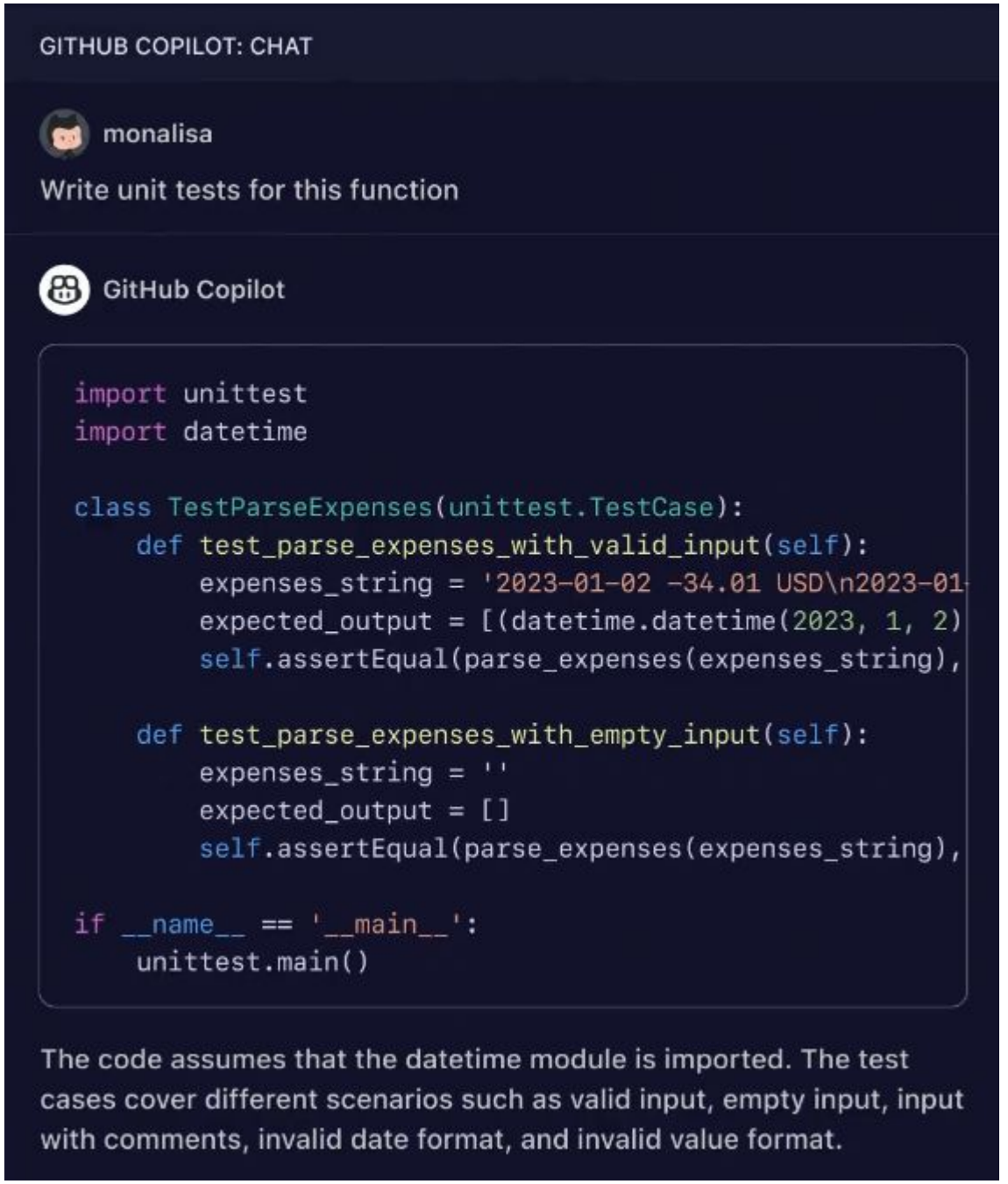


Image credit: https://github.com/features/copilot

Language Models

Education

You

Explain photosynthesis to a five-year old.

ChatGPT

Sure, I'll explain photosynthesis in a simple way for a five-year-old:

Imagine plants are like hungry little chefs. They need food to grow big and strong, just like y
need your snacks and meals.

Now, these plant chefs are very smart because they can make their own food, and they do i
the help of sunlight, air, and water. It's like a magical recipe!

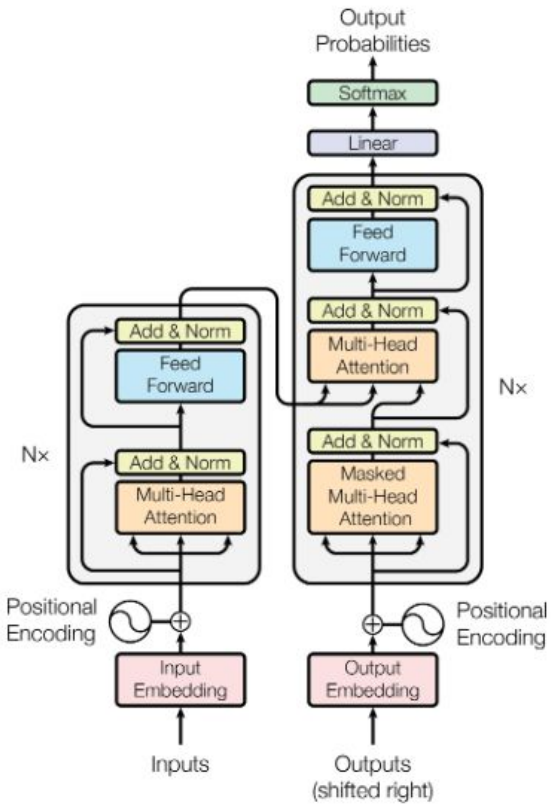
The plant uses its special leaves to catch sunlight, just like you might use a flashlight to find
toys in the dark. Then, the plant takes a big gulp of air, just like you breathe in and out. Final

Language Models

Customer Service Chatbots

Image credit: https://www.ultimate.ai/ultimategpt-chatgpt-for-customer-service

Natural Language Processing

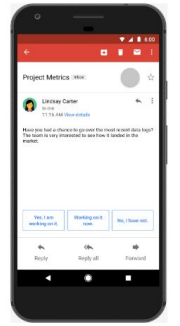


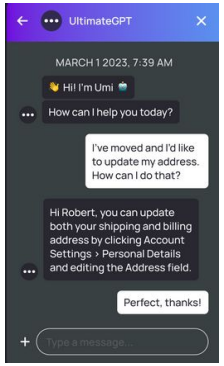
You
Explain photosynthesis to a five-year old.

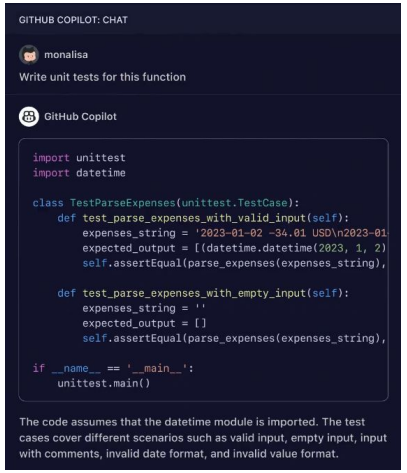
ChatGPT
Sure, I'll explain photosynthesis in a simple way for a five-year-old:

Imagine plants are like hungry little chefs. They need food to grow big and strong, just like you need your snacks and meals.

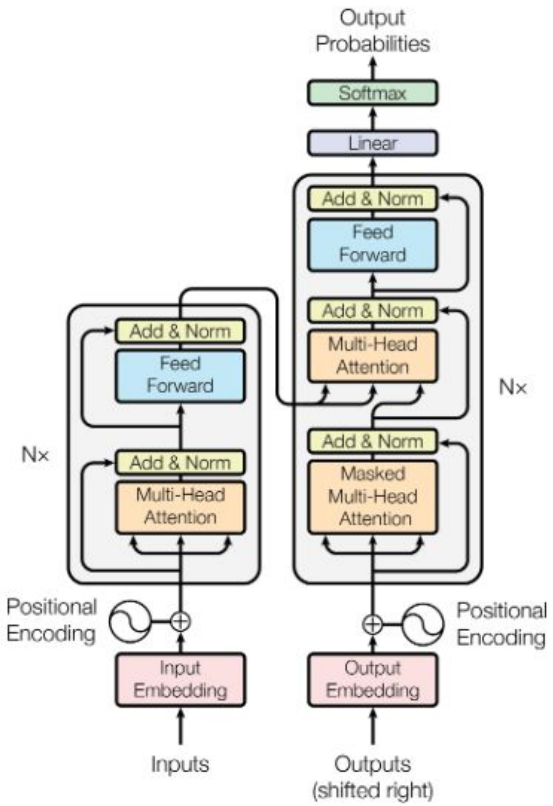
Now, these plant chefs are very smart because they can make their own food, and they do it with the help of sunlight, air, and water. It's like a magical recipe!



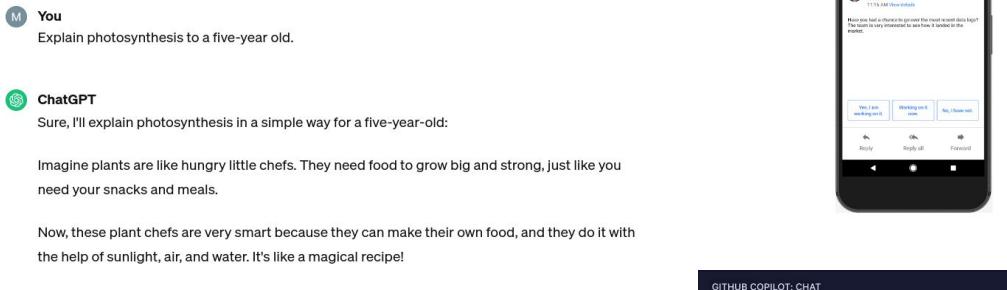


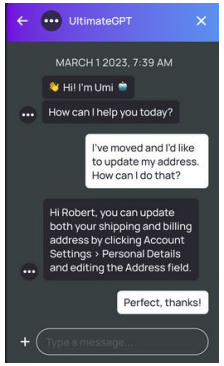


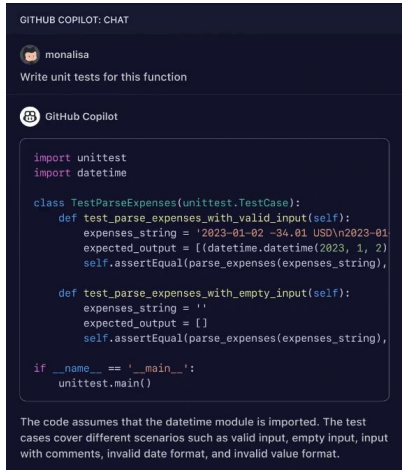
Natural Language Processing



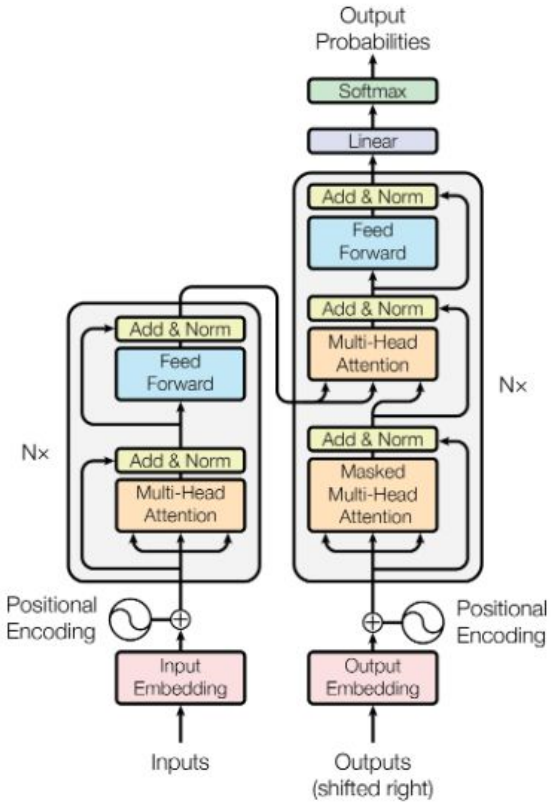
These models are black boxes.





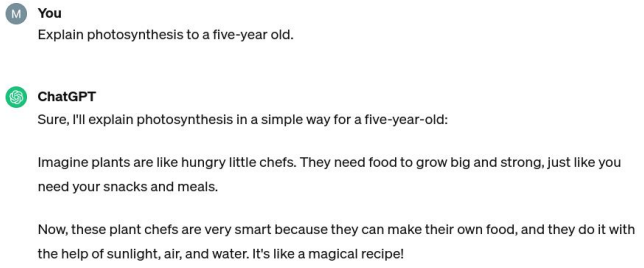


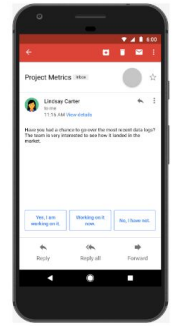
Natural Language Processing

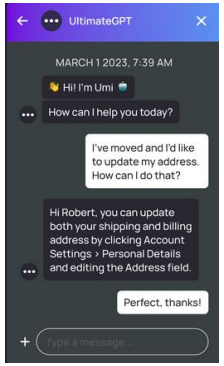


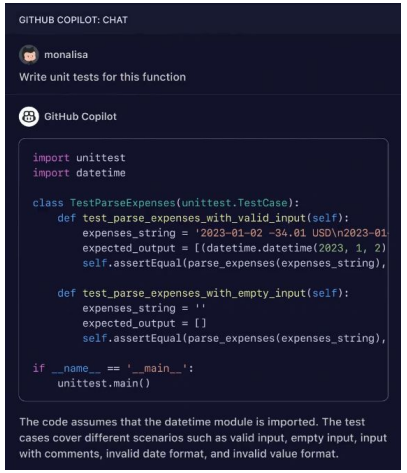
These models are black boxes.

Hard to understand their limitations & failures.

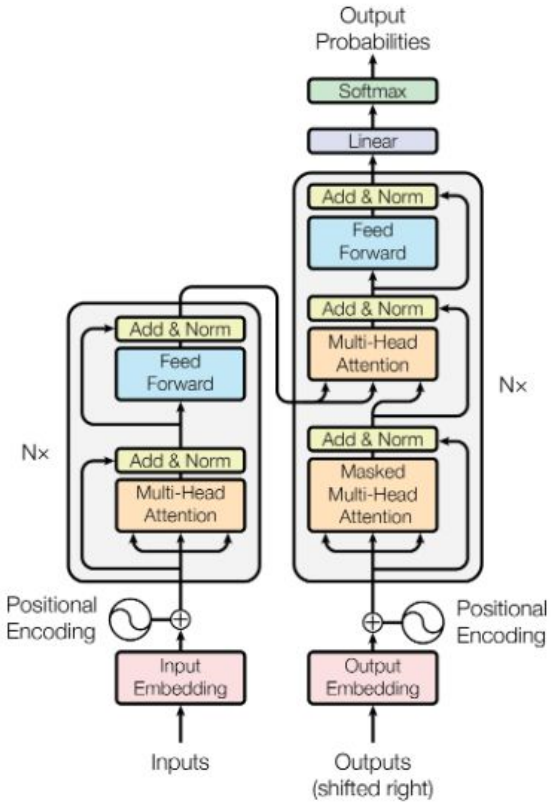






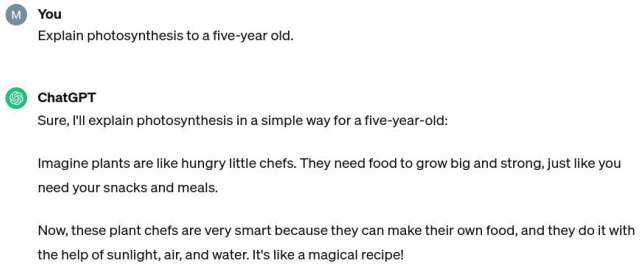


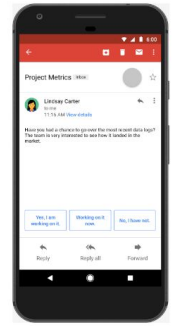
Natural Language Processing

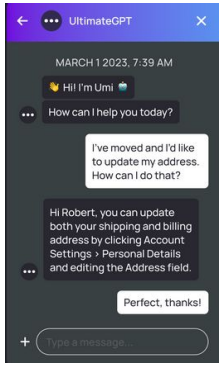


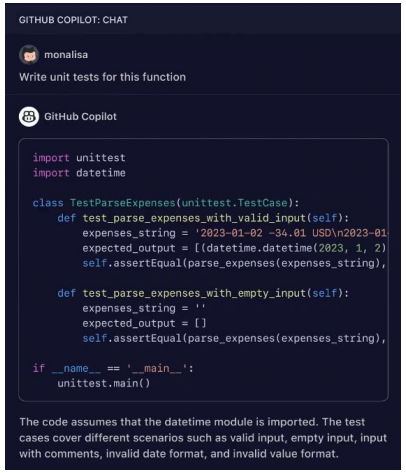
These models are black boxes.

Hard to understand their limitations & failures.









Creates problems for real-world applications.

What are the computational mechanisms needed for language understanding?

What are the computational mechanisms needed for language understanding?

What are the in-principle capabilities and limitations of neural ML architectures?

HG, in subm. 2023

HJF, TACL 2021

H**H**GLM, EMNLP 2021

HB, TACL 2021

H, TACL 2020

What are the computational mechanisms needed for language understanding?

What are the in-principle capabilities and limitations of neural ML architectures?

HG, in subm. 2023

HJF, TACL 2021

H**H**GLM, EMNLP 2021

HB, TACL 2021

H, TACL 2020

Can we use LMs to reverse-engineer human language comprehension?

HK, Cognition 2023

HFLG, PNAS 2022

HJF, PNAS 2020

HFBB, CogSci 2019

HK, EMNLP 2016

What are the computational mechanisms needed for language understanding?

What are the in-principle capabilities and limitations of neural ML architectures?

HG, in subm. 2023

HJF, TACL 2021

H**H**GLM, EMNLP 2021

HB, TACL 2021

H, TACL 2020

Can we use LMs to reverse-engineer human language comprehension?

HK, Cognition 2023

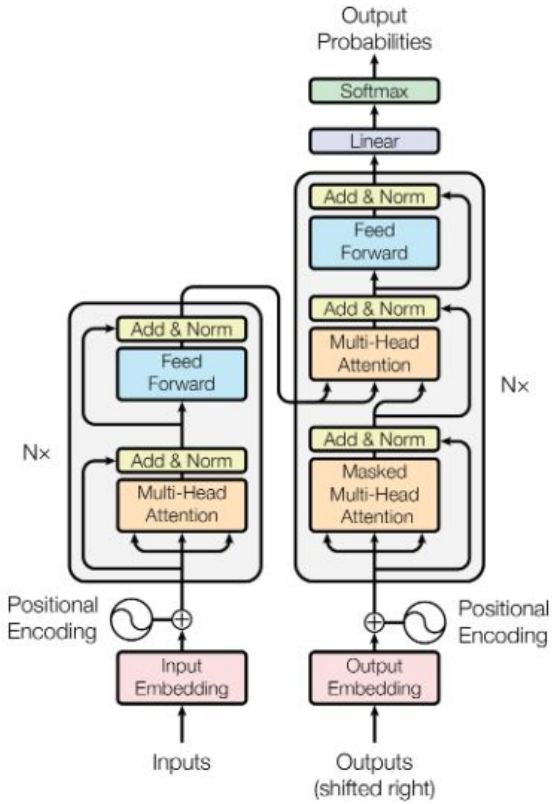
HFLG, PNAS 2022

HJF, PNAS 2020

HFBB, CogSci 2019

HK, EMNLP 2016

Capabilities and Limitations of Transformers



What capabilities can we expect, given the architecture and the training data?

Parity Function

PARITY (XOR)

Bit strings with **even number of 1s**

11
101
0000101011

1
01
000001011

Parity Function

Relevance to Language: Negation

PARITY (XOR)

Bit strings with **even number of 1s**

11	1
101	01
0000101011	000001011

Parity Function

Relevance to Language: Negation

PARITY (XOR)

Bit strings with **even number of 1s**

11
101
0000101011

1
01
000001011

happy

happy

even

Parity Function

PARITY (XOR)

Bit strings with **even number of 1s**

11
101
0000101011

1
01
000001011

Relevance to Language: Negation

happy — *even*

happy

unhappy — *odd*

not (happy)

Parity Function

PARITY (XOR)

Bit strings with **even number of 1s**

11
101
0000101011

1
01
000001011

Relevance to Language: Negation

happy — *even*

happy

unhappy — *odd*

not (happy)

not unhappy — *even*

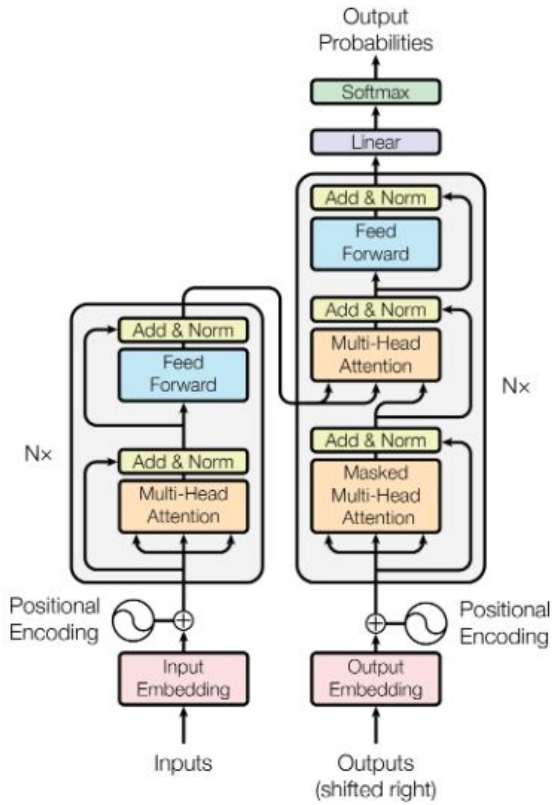
not (not (happy))

Transformers

Theorem Hahn, TACL 2020

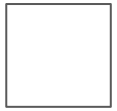
No transformer can represent PARITY robustly at all input lengths.

Transformers



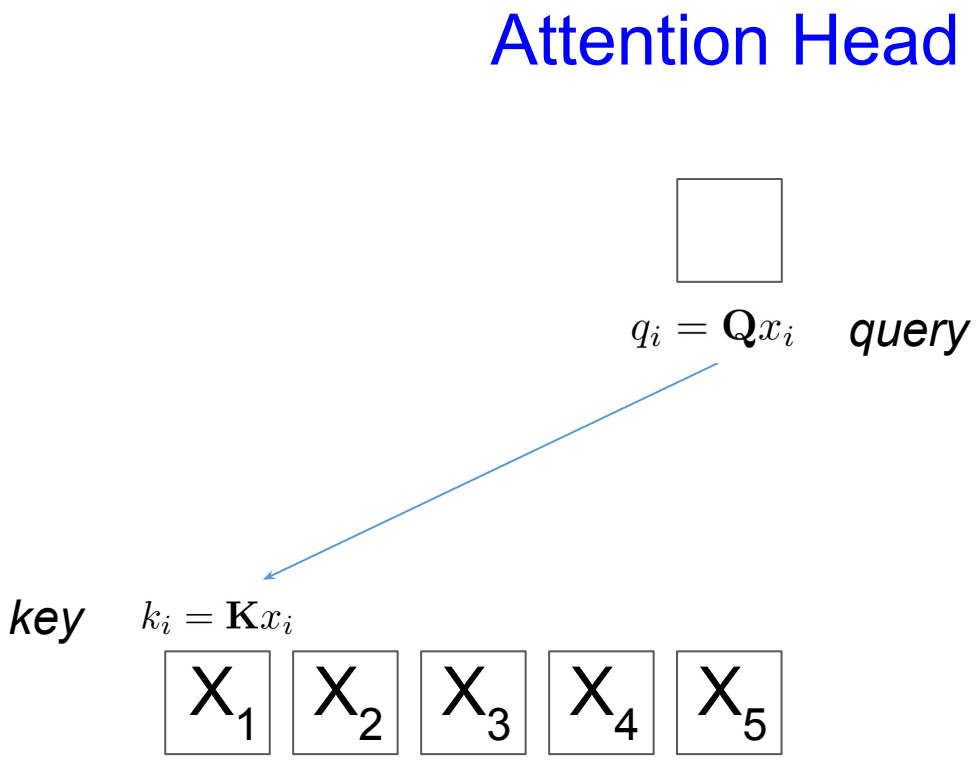
Transformers

Attention Head

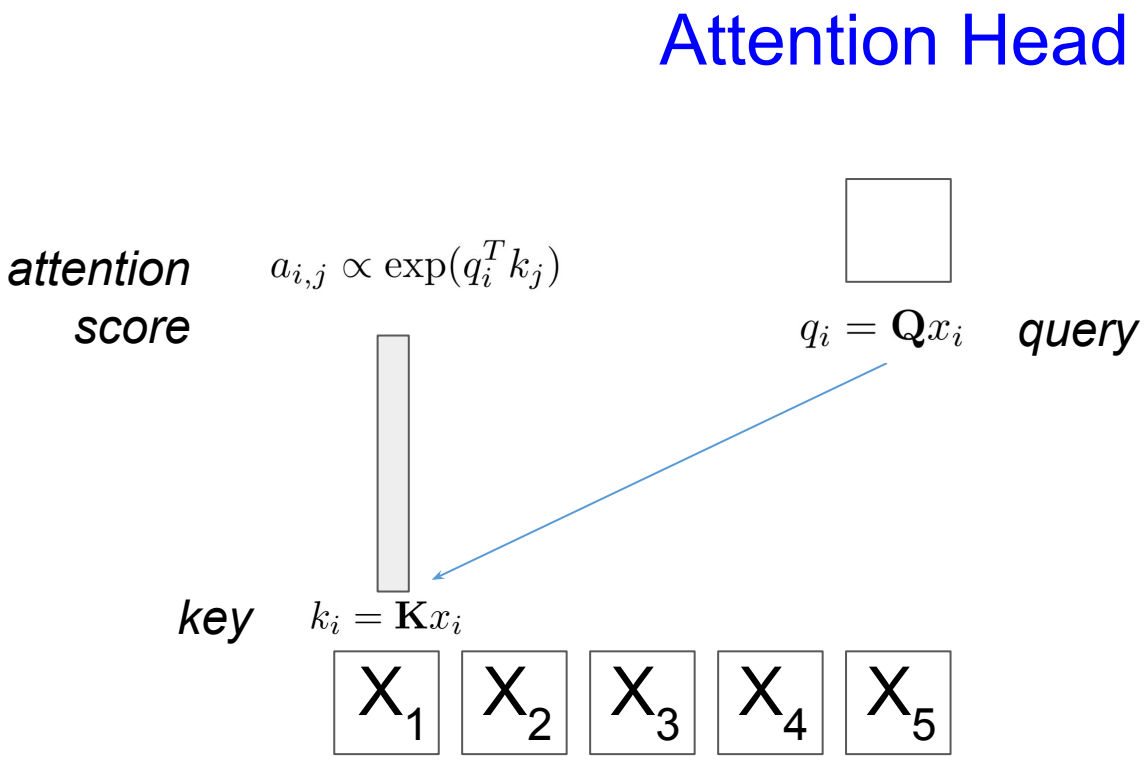


$q_i = \mathbf{Q}x_i$ *query*

Transformers

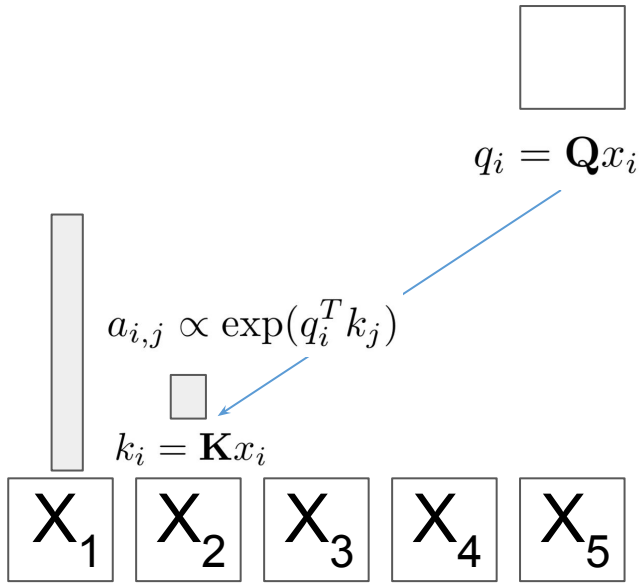


Transformers



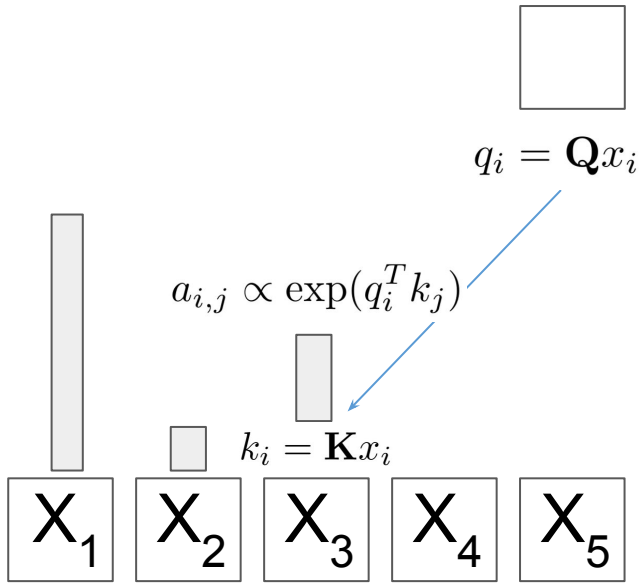
Transformers

Attention Head



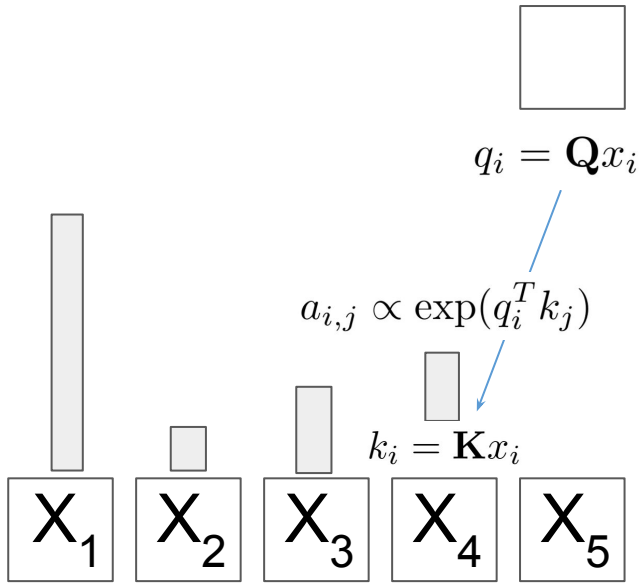
Transformers

Attention Head



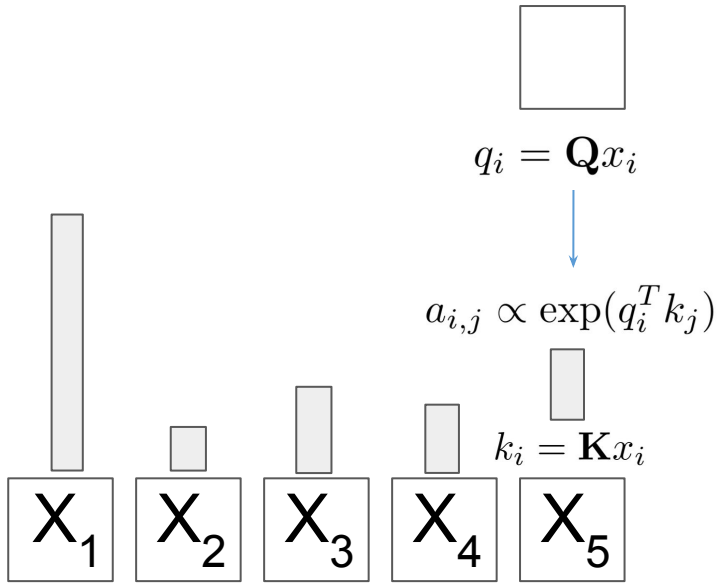
Transformers

Attention Head

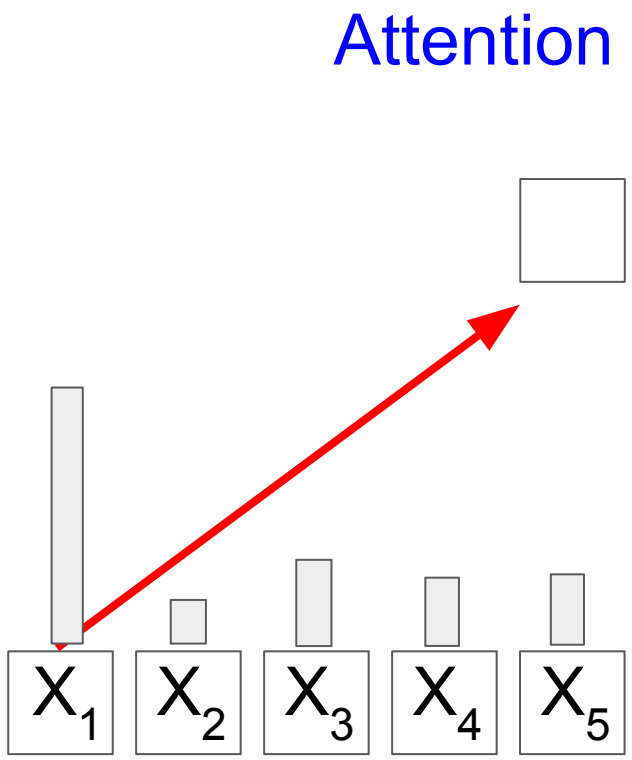


Transformers

Attention Head

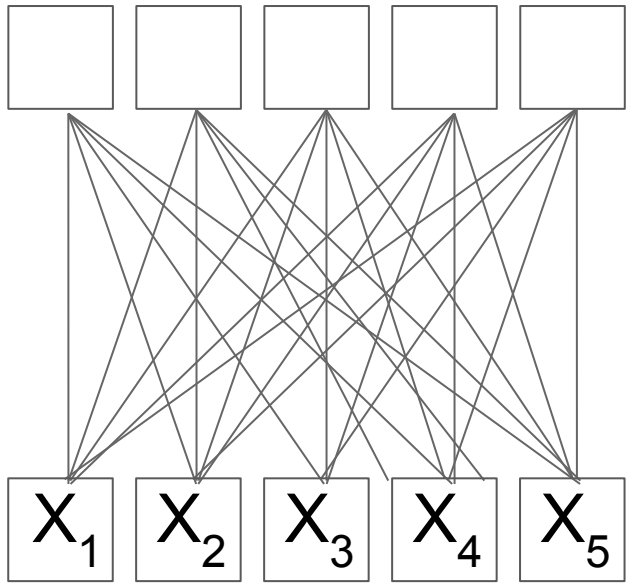


Transformers

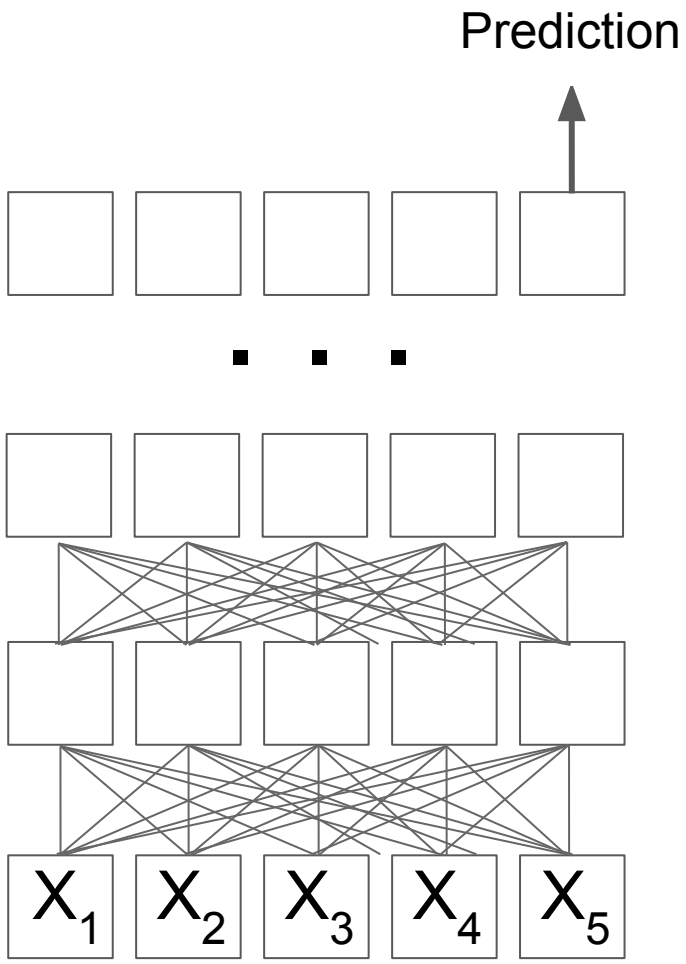


Here: “hard attention”. Similar result for “soft attention”.

Transformers

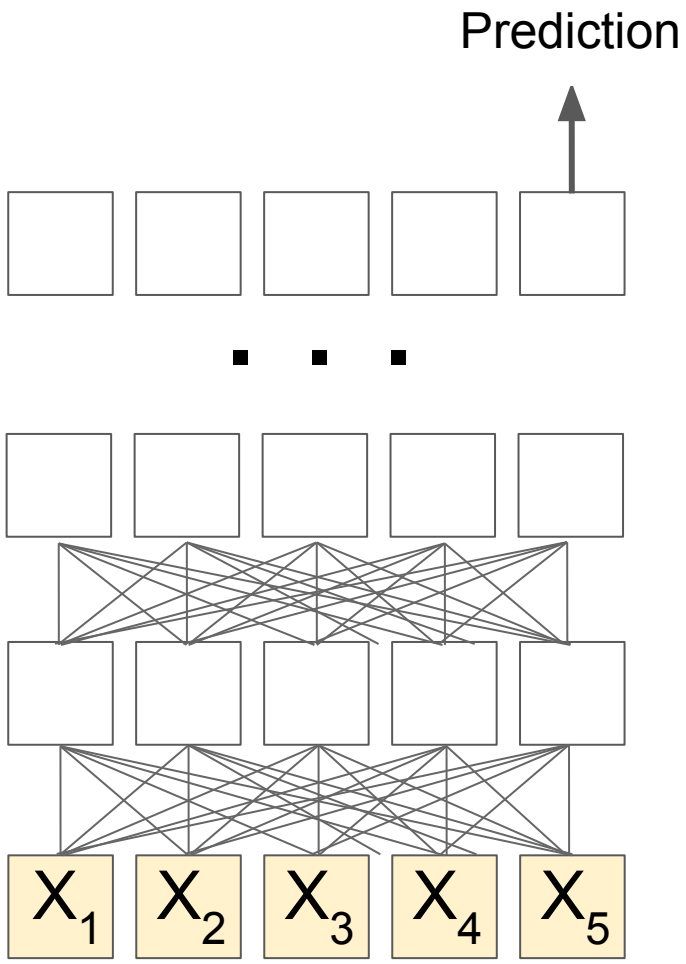


Transformers



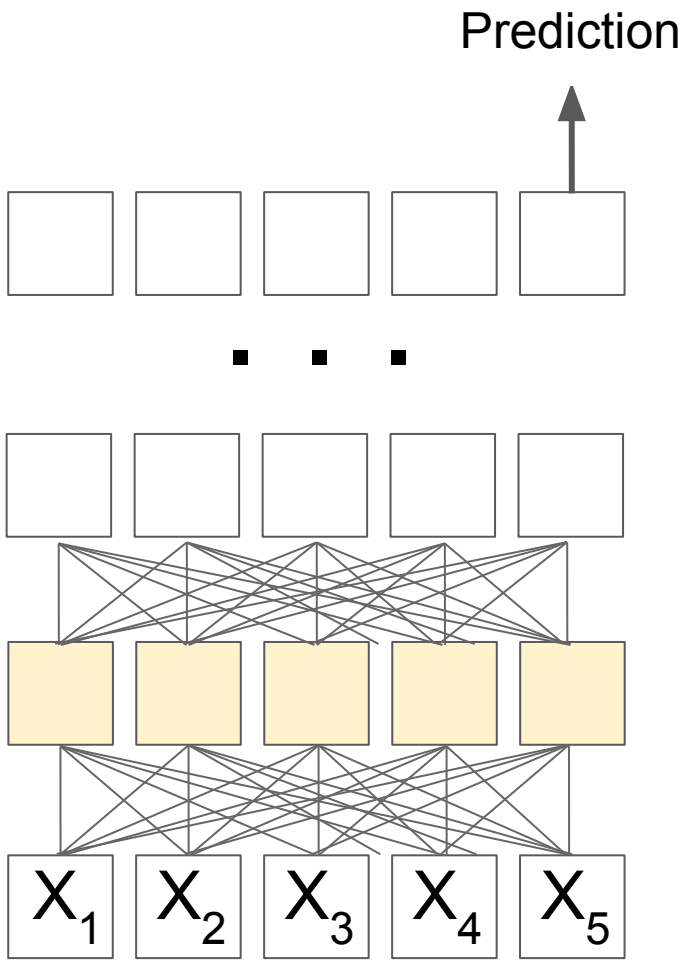
Transformers

Information is
propagated upwards



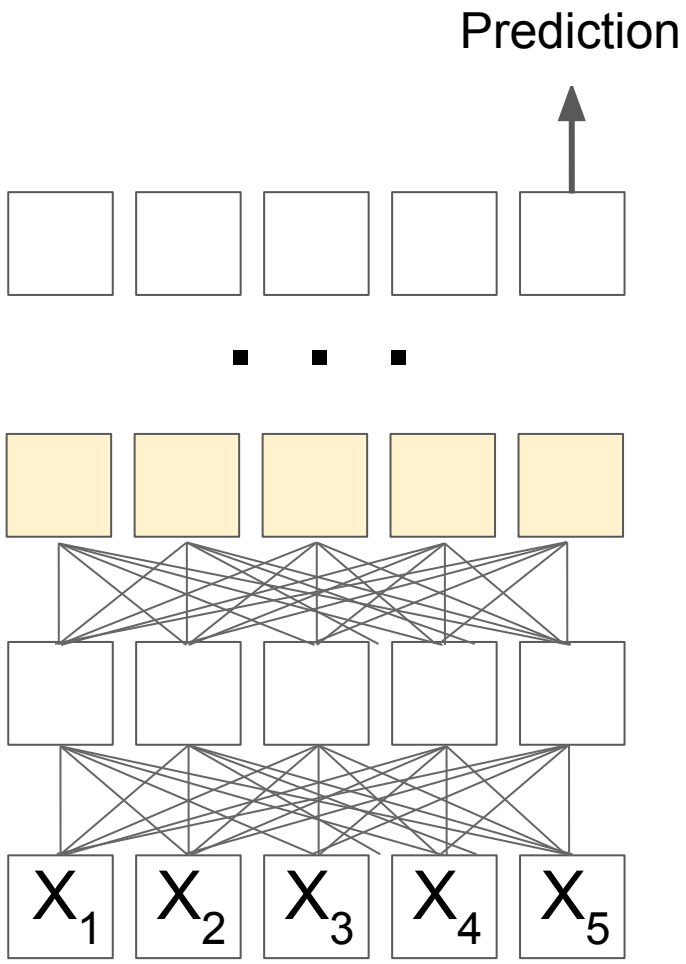
Transformers

Information is
propagated upwards



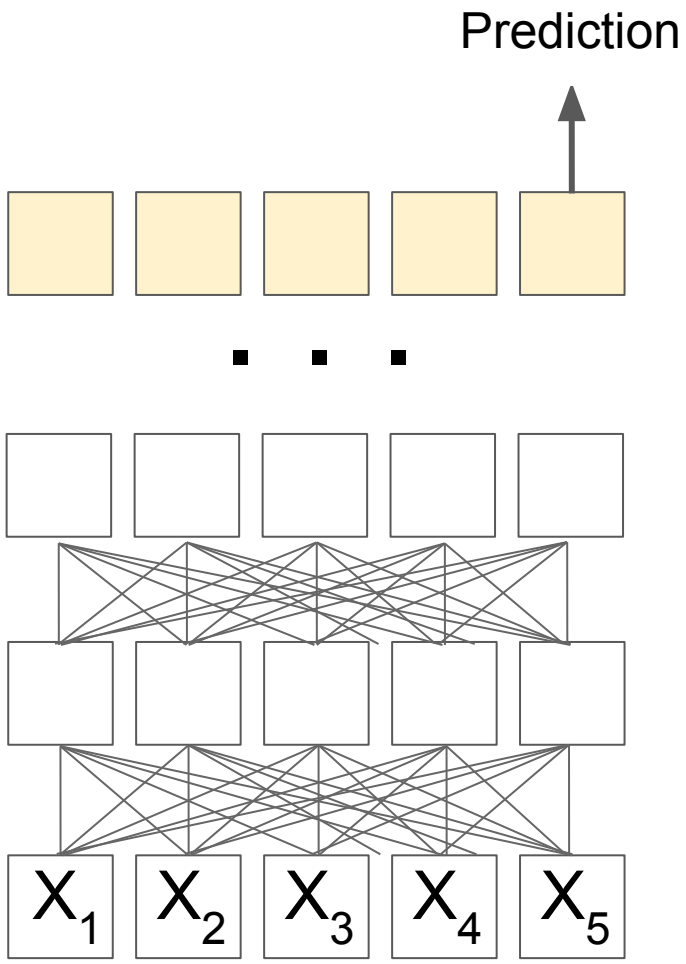
Transformers

Information is
propagated upwards



Transformers

Information is
propagated upwards

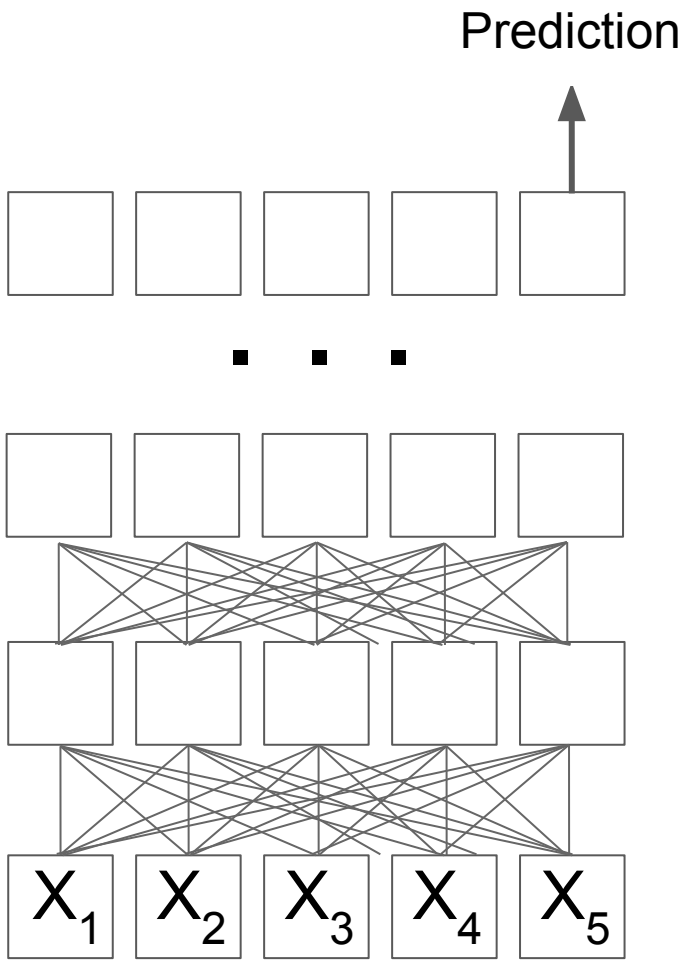


Transformers

Information is propagated upwards

Suspected to have limited expressive capacity

(Tran et al., 2018; Dehghani et al., 2019; Shen et al., 2018; Chen et al., 2018; Hao et al., 2019)

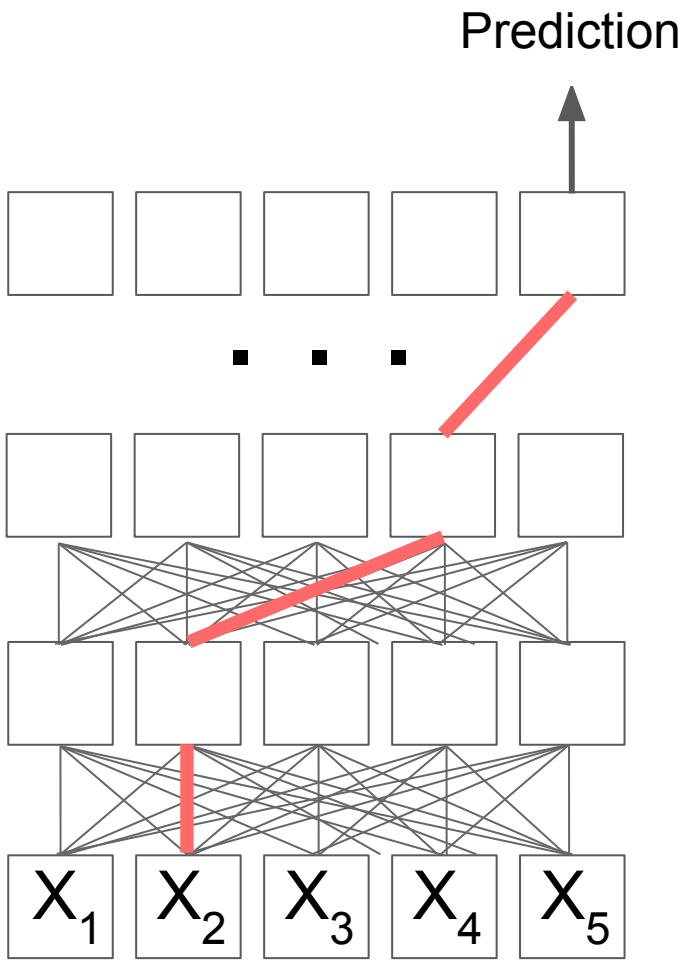


Transformers

Information is propagated upwards

Suspected to have limited expressive capacity

(Tran et al., 2018; Dehghani et al., 2019; Shen et al., 2018; Chen et al., 2018; Hao et al., 2019)



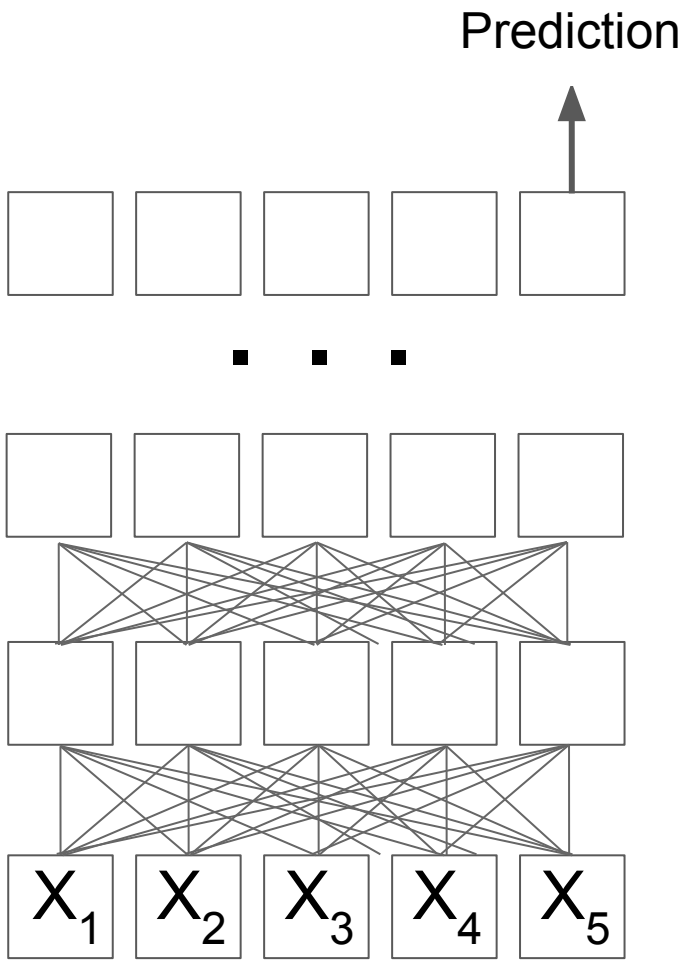
Computation paths have bounded length

Transformers

Information is propagated upwards

Suspected to have limited expressive capacity

But, previously no proof for this intuition



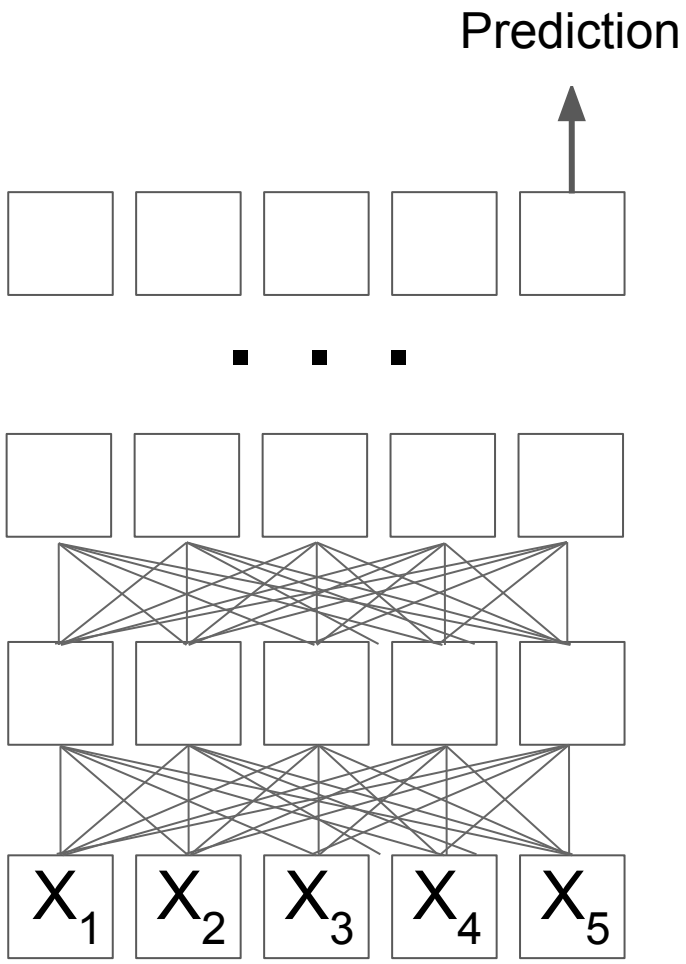
Computation paths have bounded length

Transformers

Information is propagated upwards

Suspected to have limited expressive capacity

But, previously no proof for this intuition

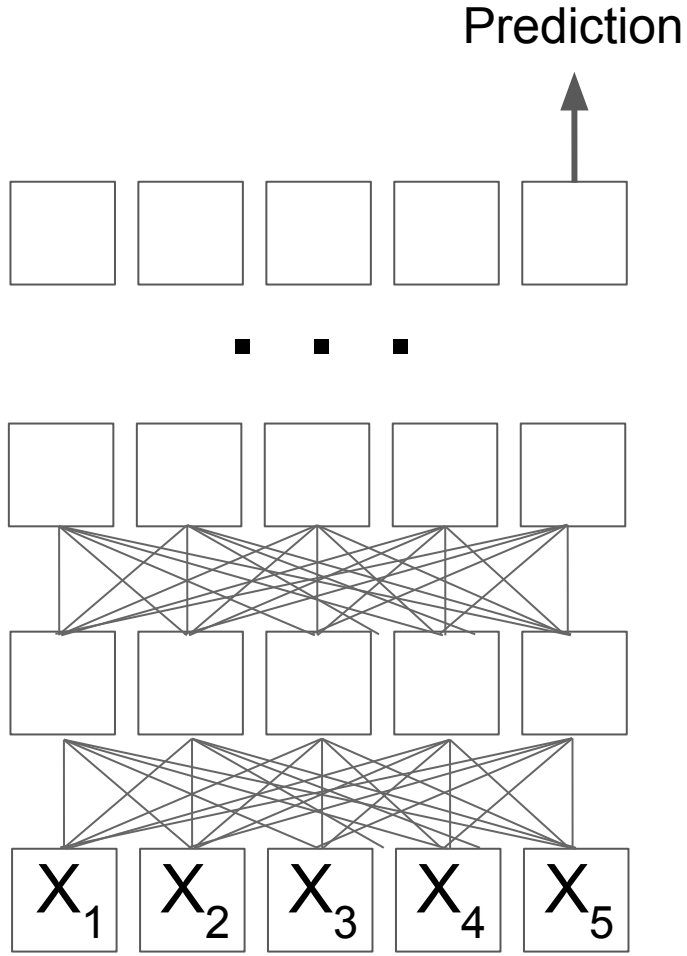


Theorem (TACL 2020)

No transformer can represent PARITY robustly at all input lengths.

Proof Strategy

Assume we have a transformer for PARITY.

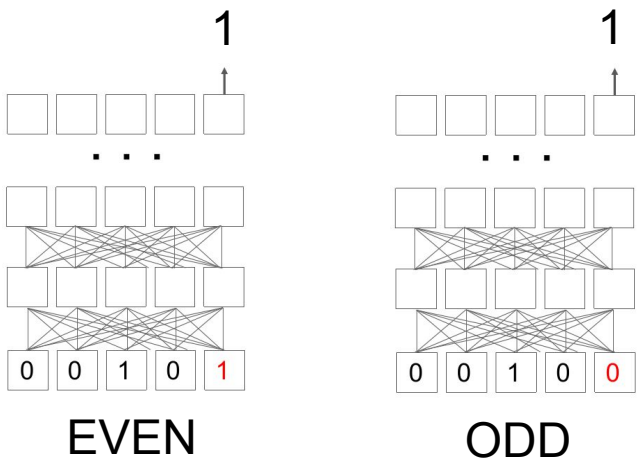


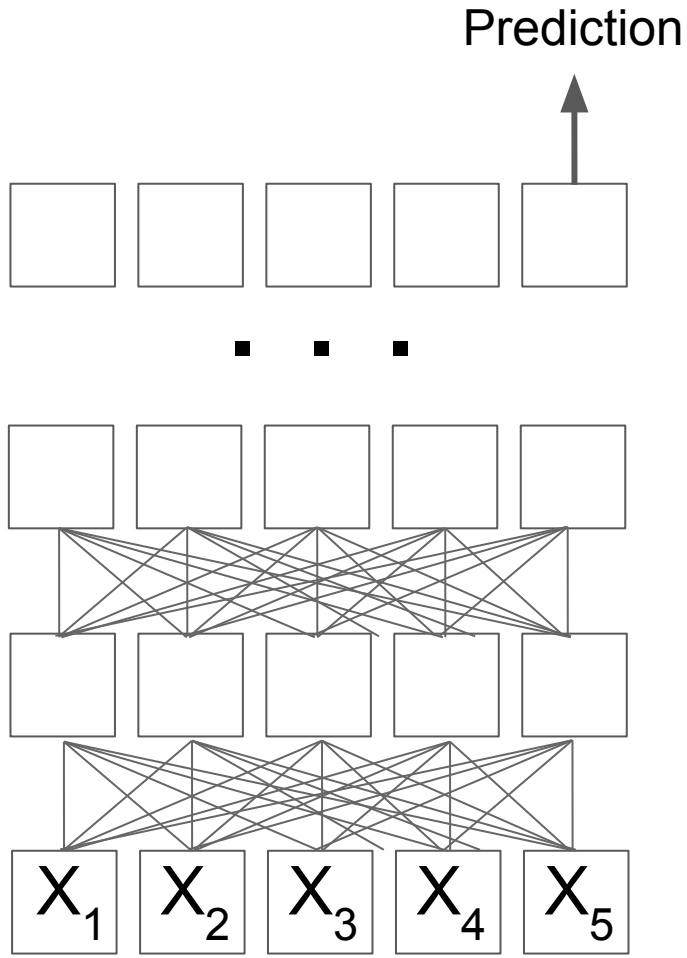
Hahn, TACL 2020

Proof Strategy

Assume we have a transformer for PARITY.

Construct two inputs that are classified the same



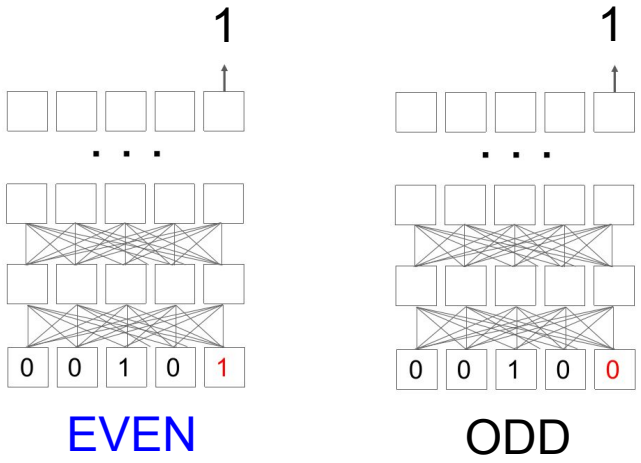


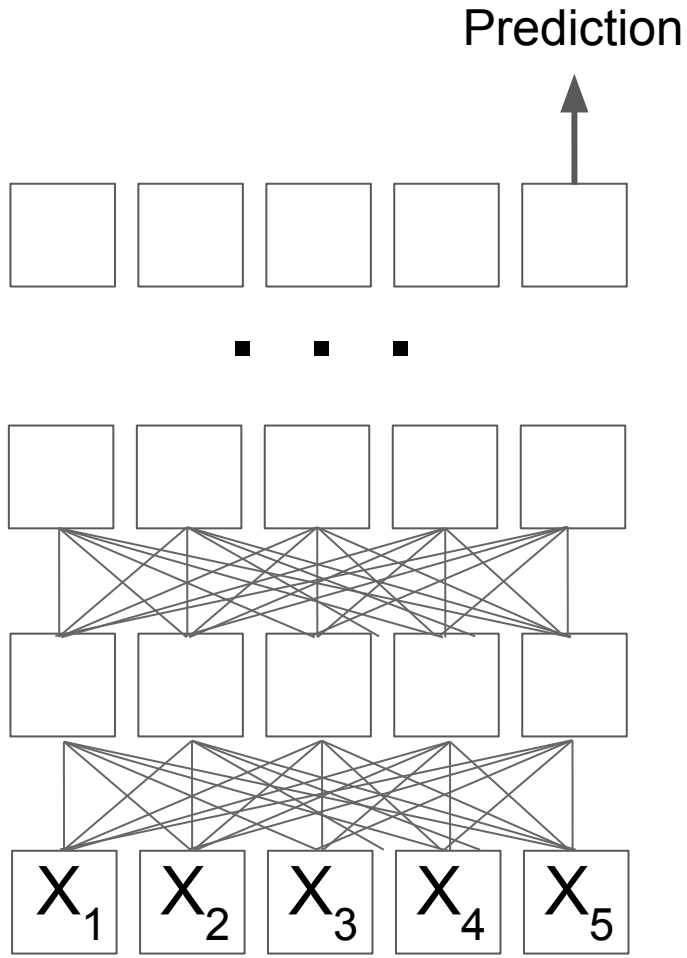
Hahn, TACL 2020

Proof Strategy

Assume we have a transformer for PARITY.

Construct two inputs that are classified the same, even though one is EVEN and the other is ODD.



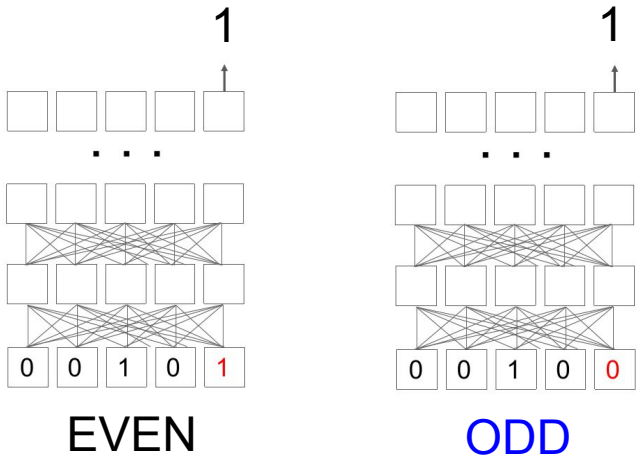


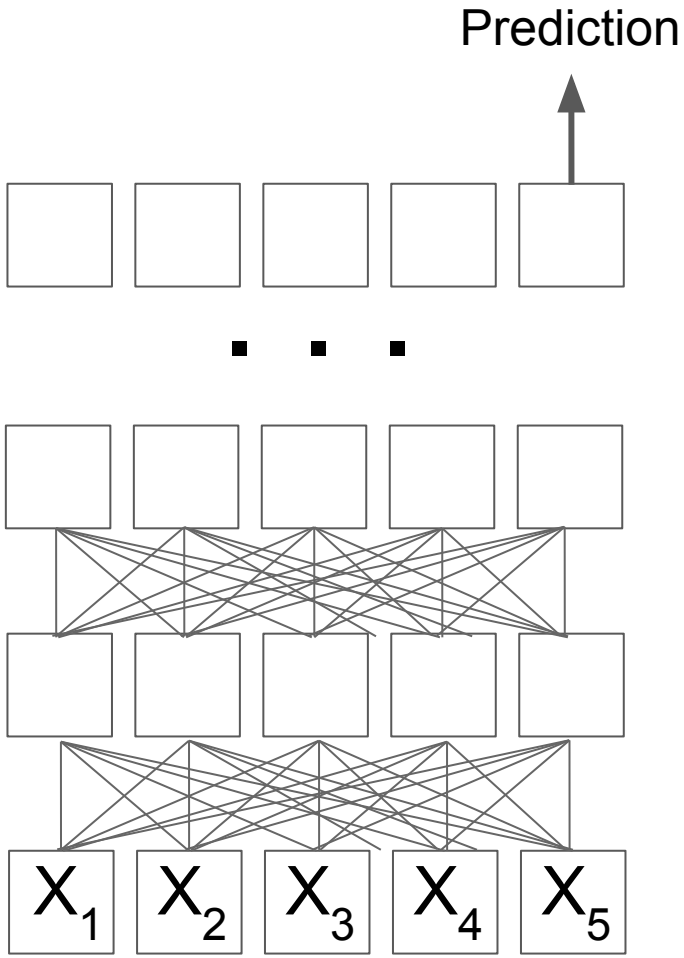
Hahn, TACL 2020

Proof Strategy

Assume we have a transformer for PARITY.

Construct two inputs that are classified the same, even though one is EVEN and the other is ODD.





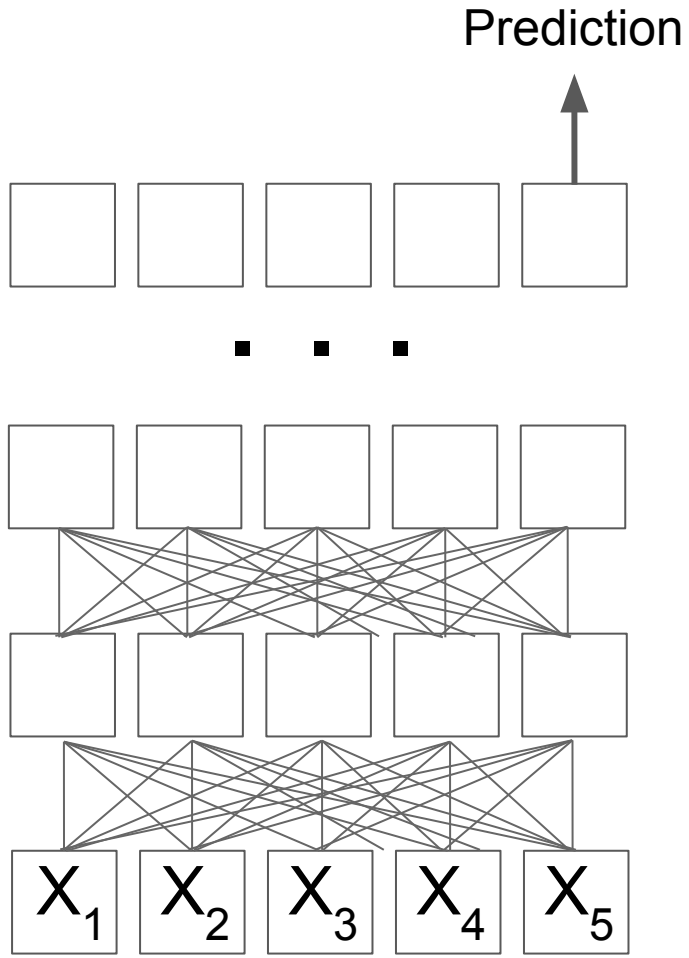
Hahn, TACL 2020

Proof Strategy

Assume we have a transformer for PARITY.

Construct two inputs that are classified the same, even though one is EVEN and the other is ODD.

Method: Strategically fix a few input bits to 'distract' the transformer, so that it ignores part of the input.



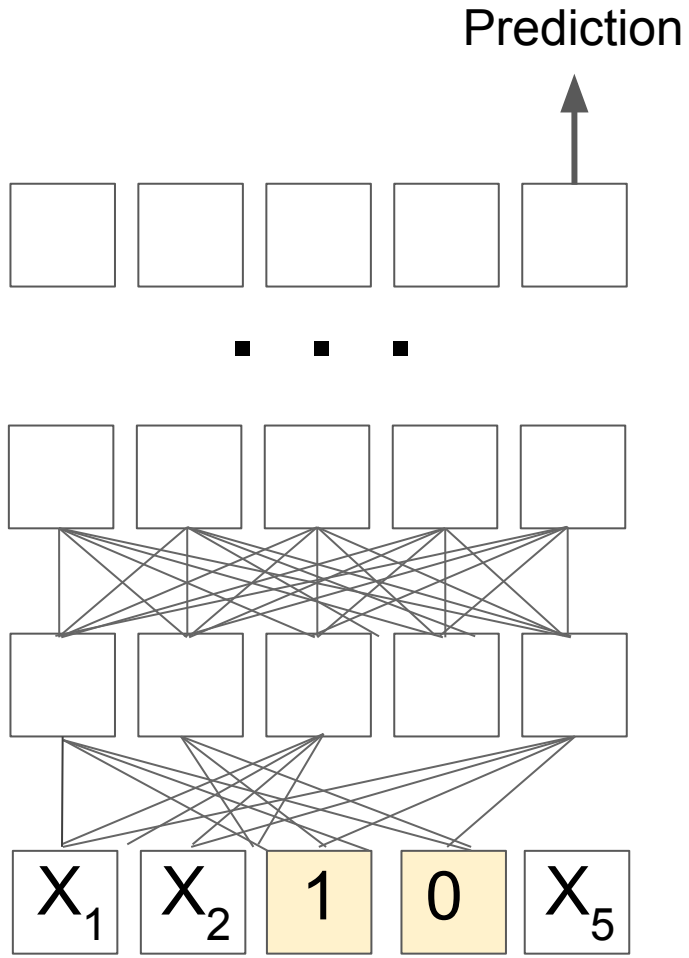
Hahn, TACL 2020

Proof Strategy

Assume we have a transformer for PARITY.

Construct two inputs that are classified the same, even though one is EVEN and the other is ODD.

Method: Strategically fix a few input bits to 'distract' the transformer, so that it ignores part of the input.



Hahn, TACL 2020

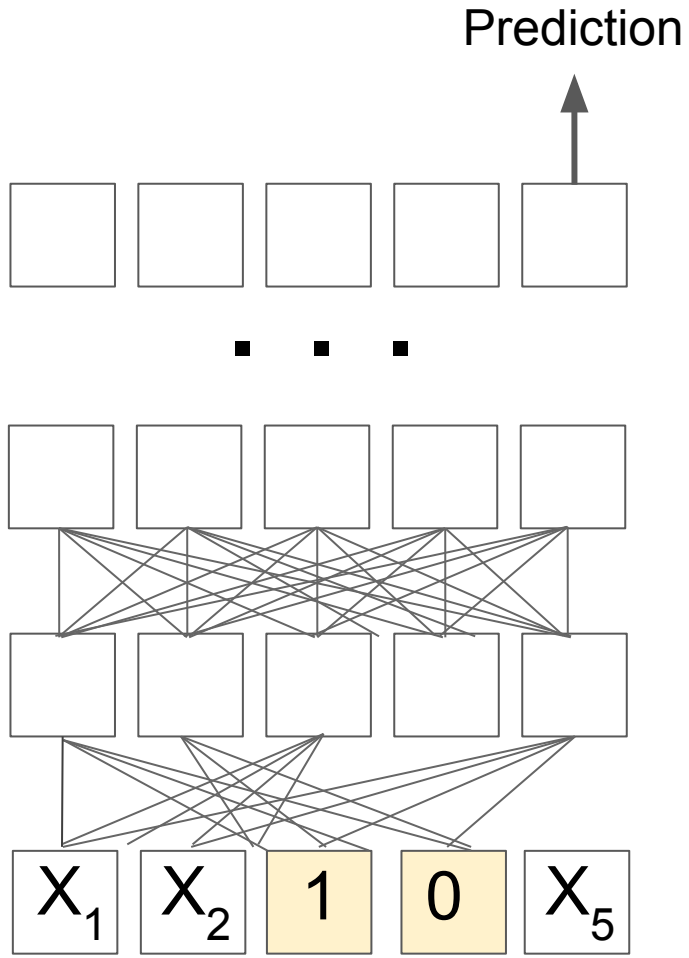
Proof Strategy

Assume we have a transformer for PARITY.

Construct two inputs that are classified the same, even though one is EVEN and the other is ODD.

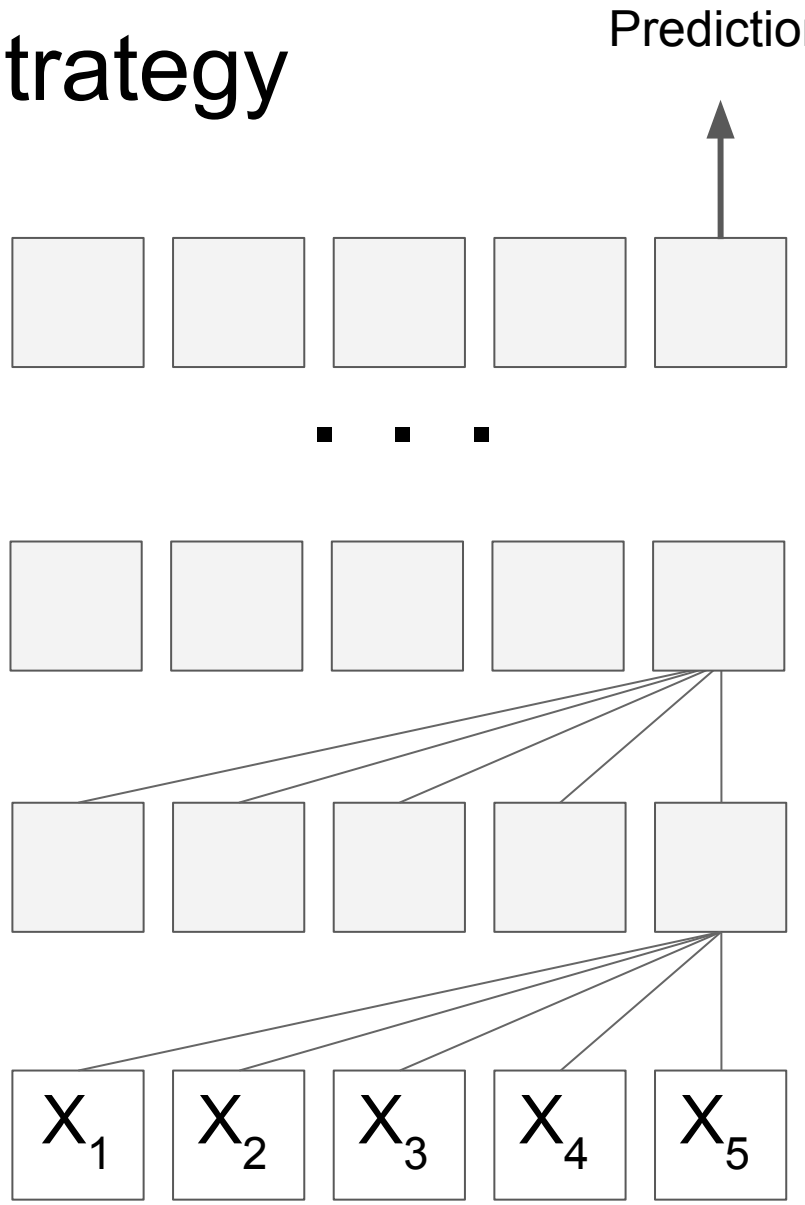
Method: Strategically fix a few input bits to 'distract' the transformer, so that it ignores part of the input.

Approach: Probabilistic Method



Hahn, TACL 2020

Proof Strategy

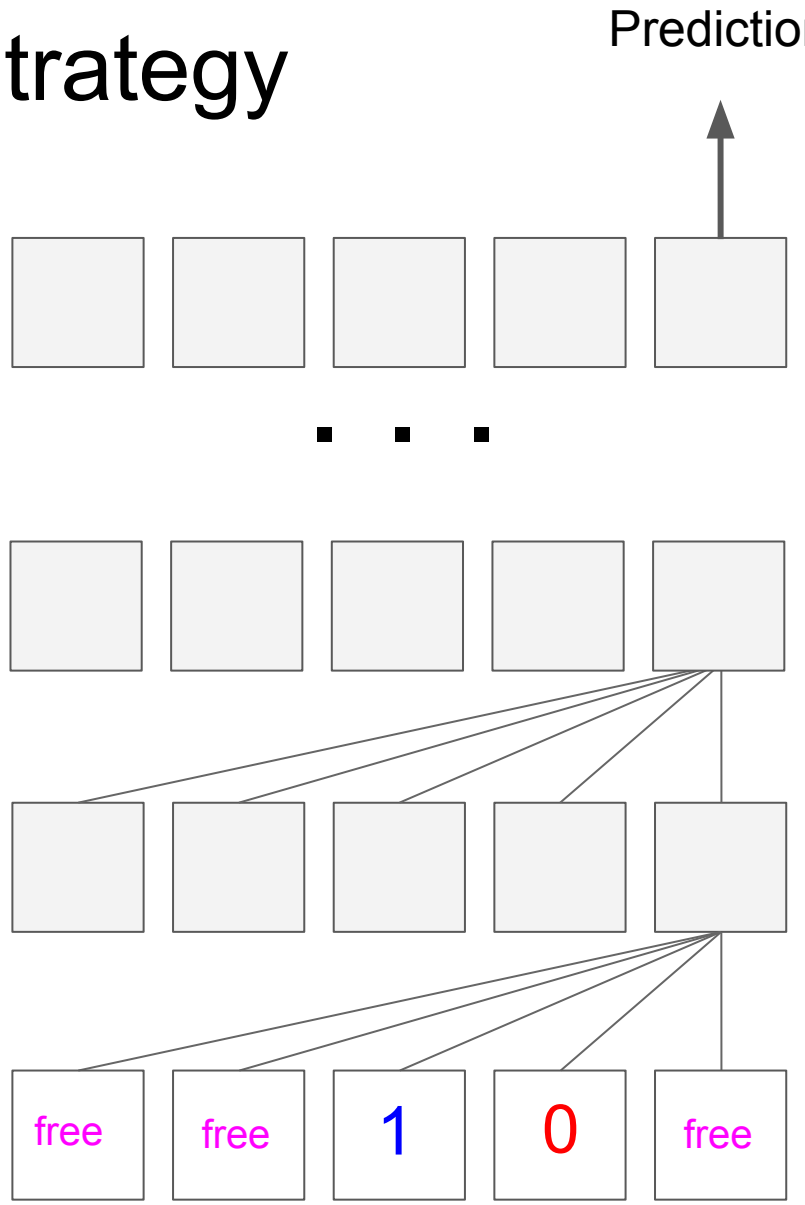


Set each input i.i.d. to

free	with p=80%
0	with p=10%
1	with p=10%

Hahn, TACL 2020

Proof Strategy

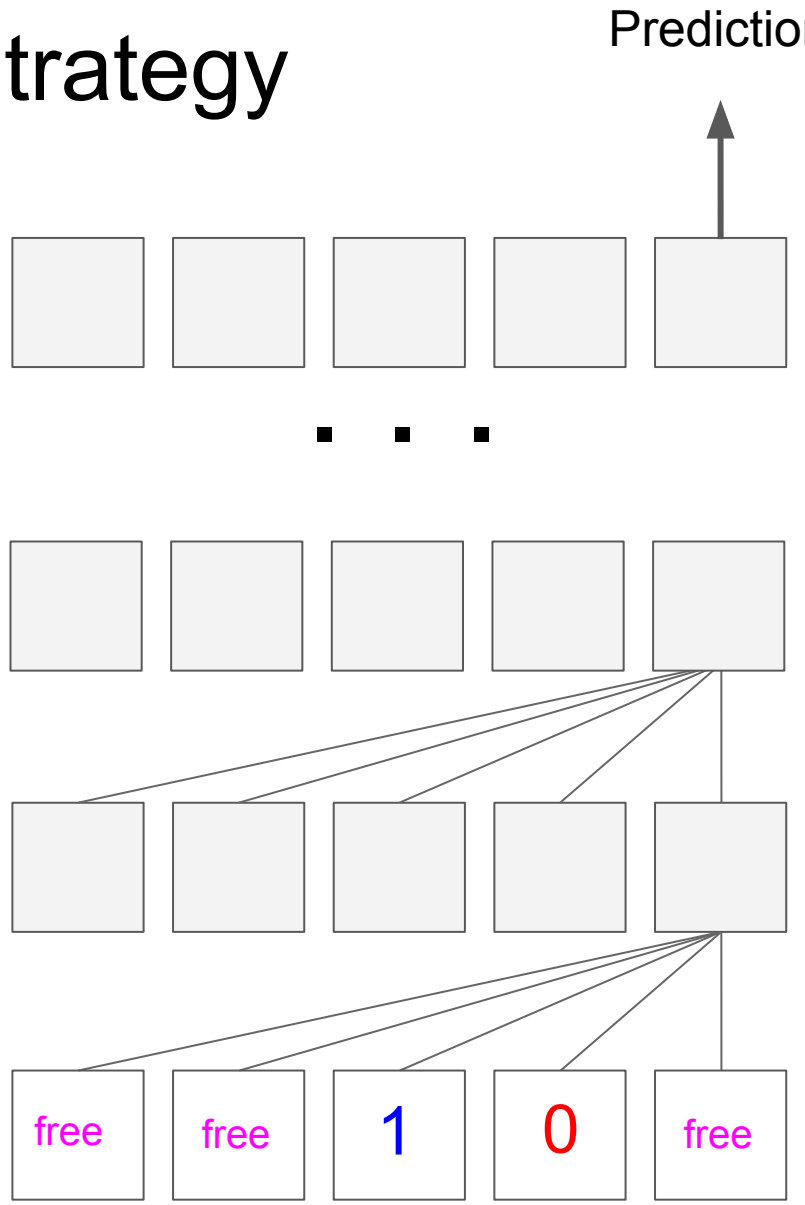


Set each input i.i.d. to

free	with p=80%
0	with p=10%
1	with p=10%

Hahn, TACL 2020

Proof Strategy



Set each input i.i.d. to

free	with p=80%
0	with p=10%
1	with p=10%

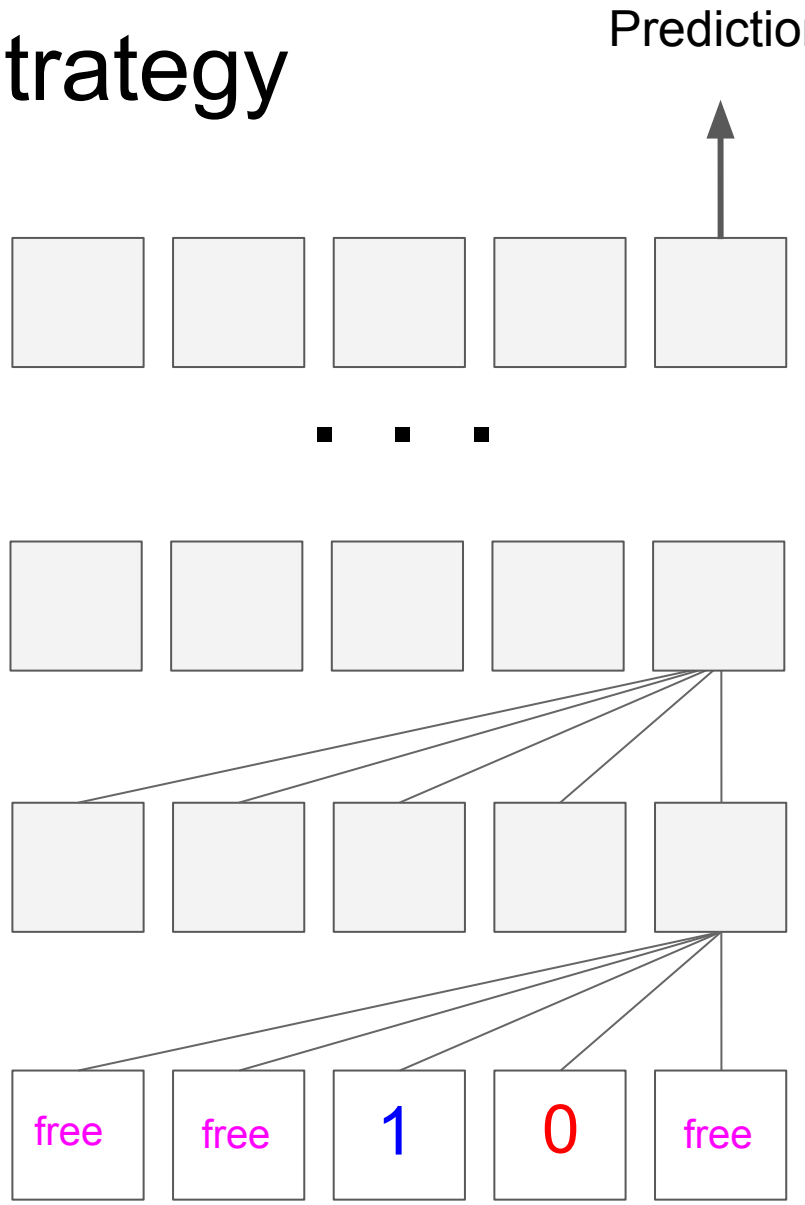
Analogous to Random Restrictions from Circuit Complexity!

Furst, Saxe, Sipser 1984

Håstad 1986

Hahn, TACL 2020

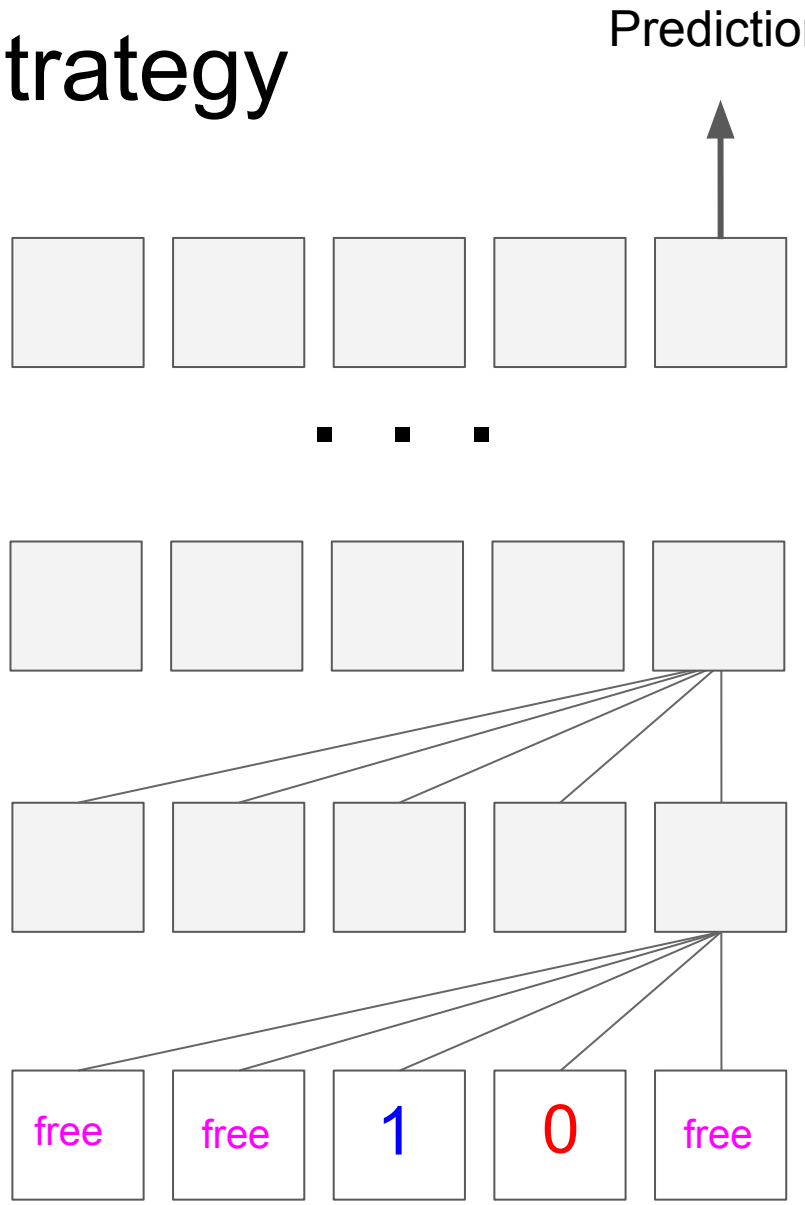
Proof Strategy



What is probability that the model ignores some free inputs?

Hahn, TACL 2020

Proof Strategy

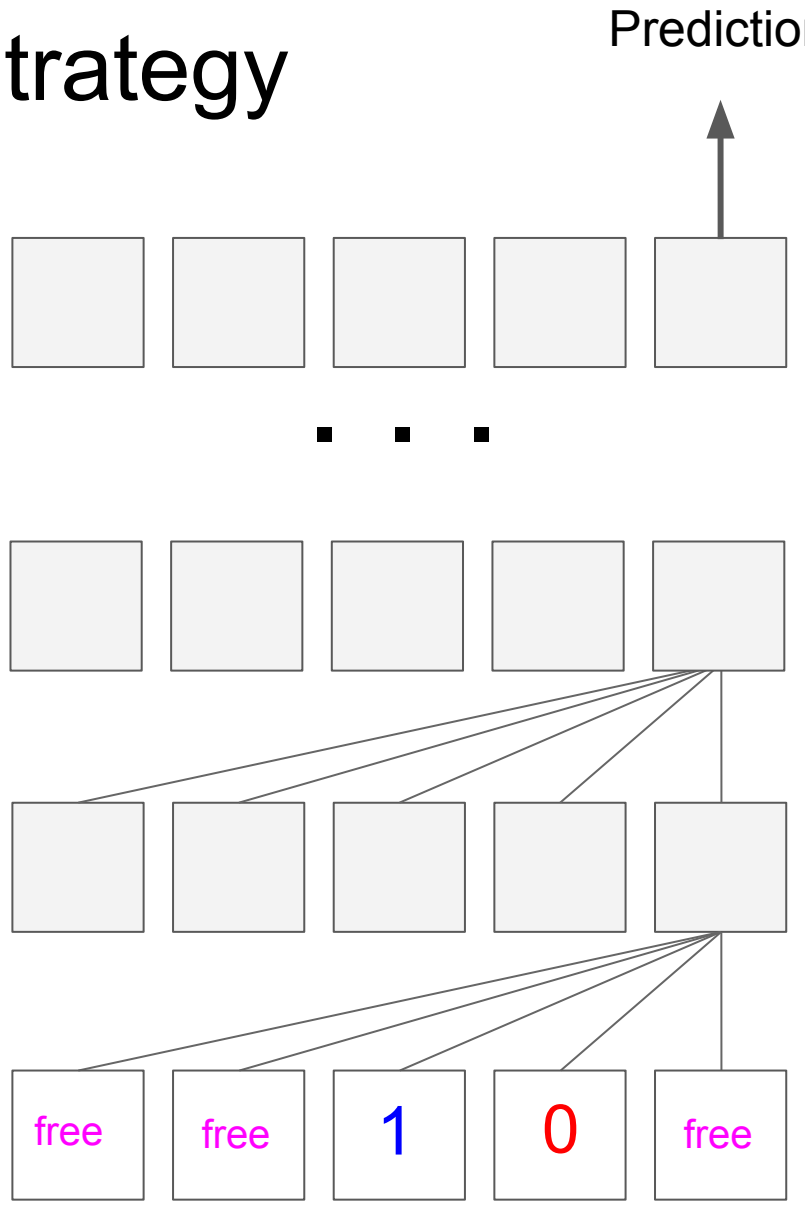


What is probability that the model ignores some free inputs?

Enough to show that this is > 0 if the input string is sufficiently long!

Hahn, TACL 2020

Proof Strategy



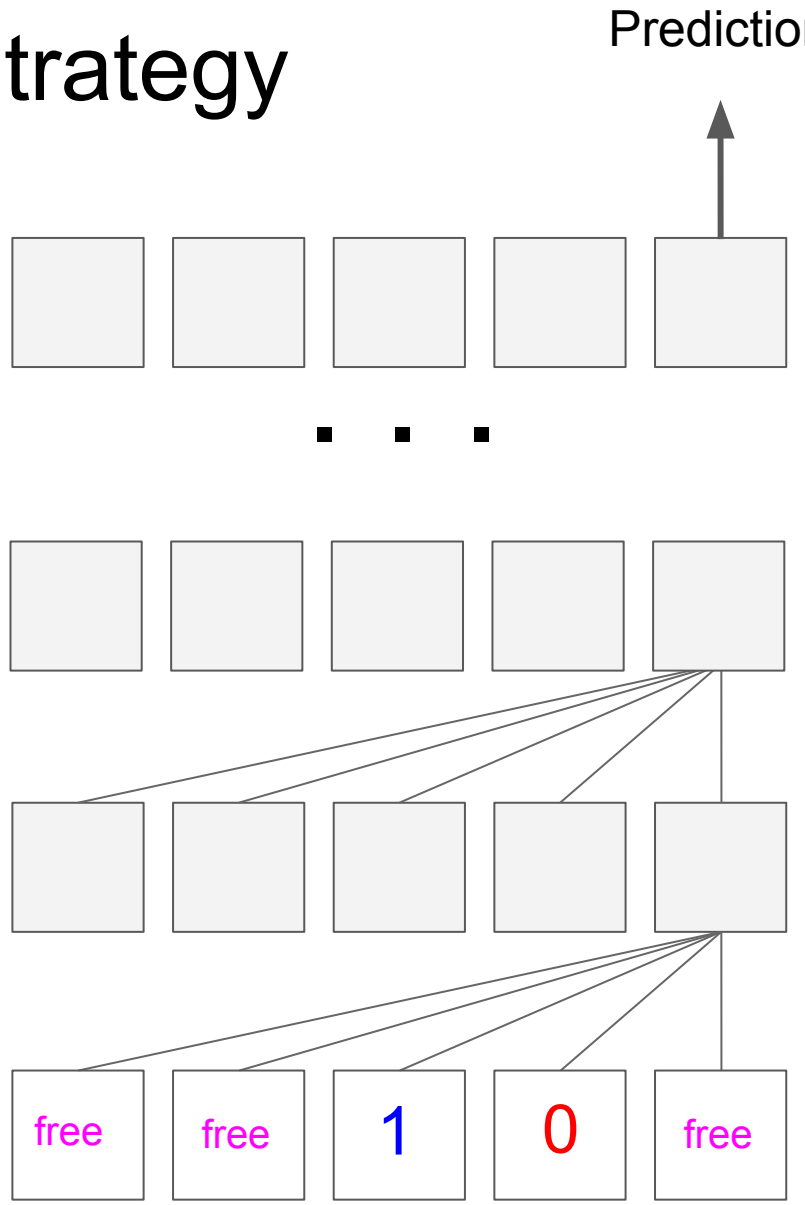
What is probability that the model ignores some free inputs?

Enough to show that this is > 0 if the input string is sufficiently long!

Show this by

Hahn, TACL 2020

Proof Strategy



What is probability that the model ignores some free inputs?

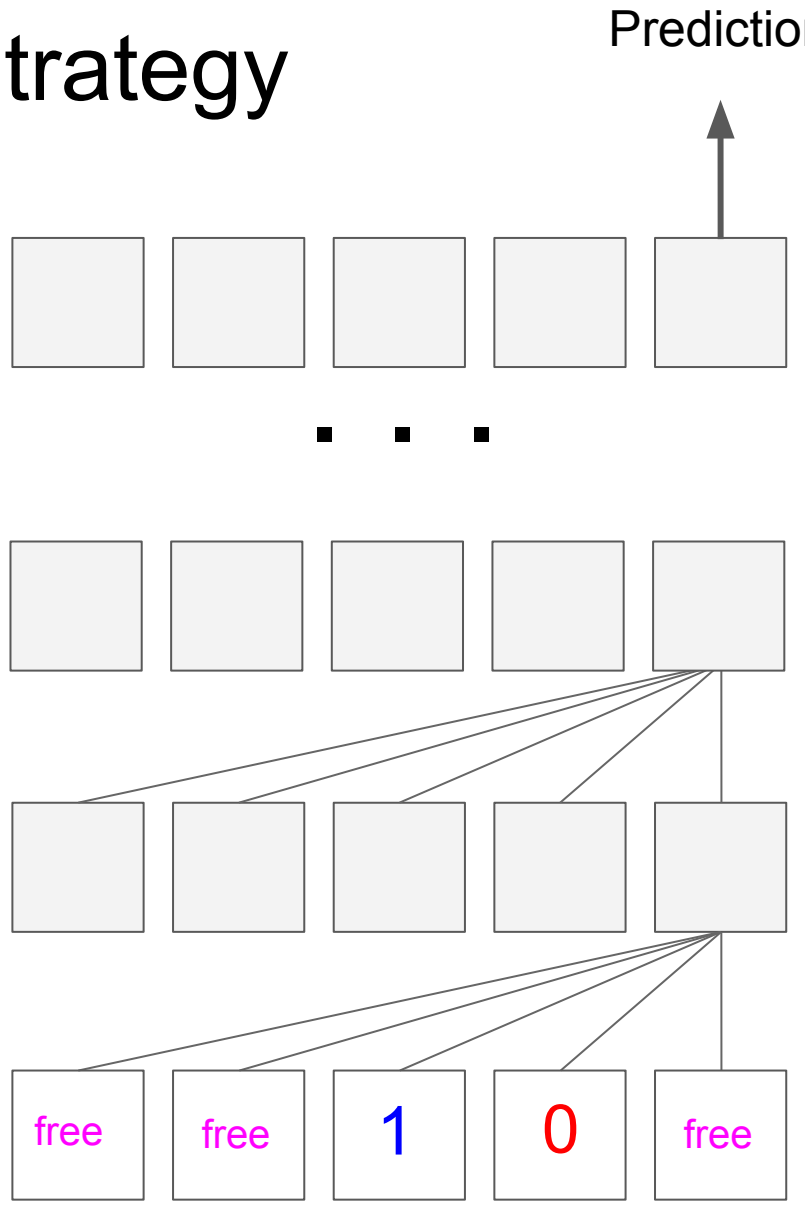
Enough to show that this is > 0 if the input string is sufficiently long!

Show this by

1. calculating for each head

Hahn, TACL 2020

Proof Strategy



What is probability that the model ignores some free inputs?

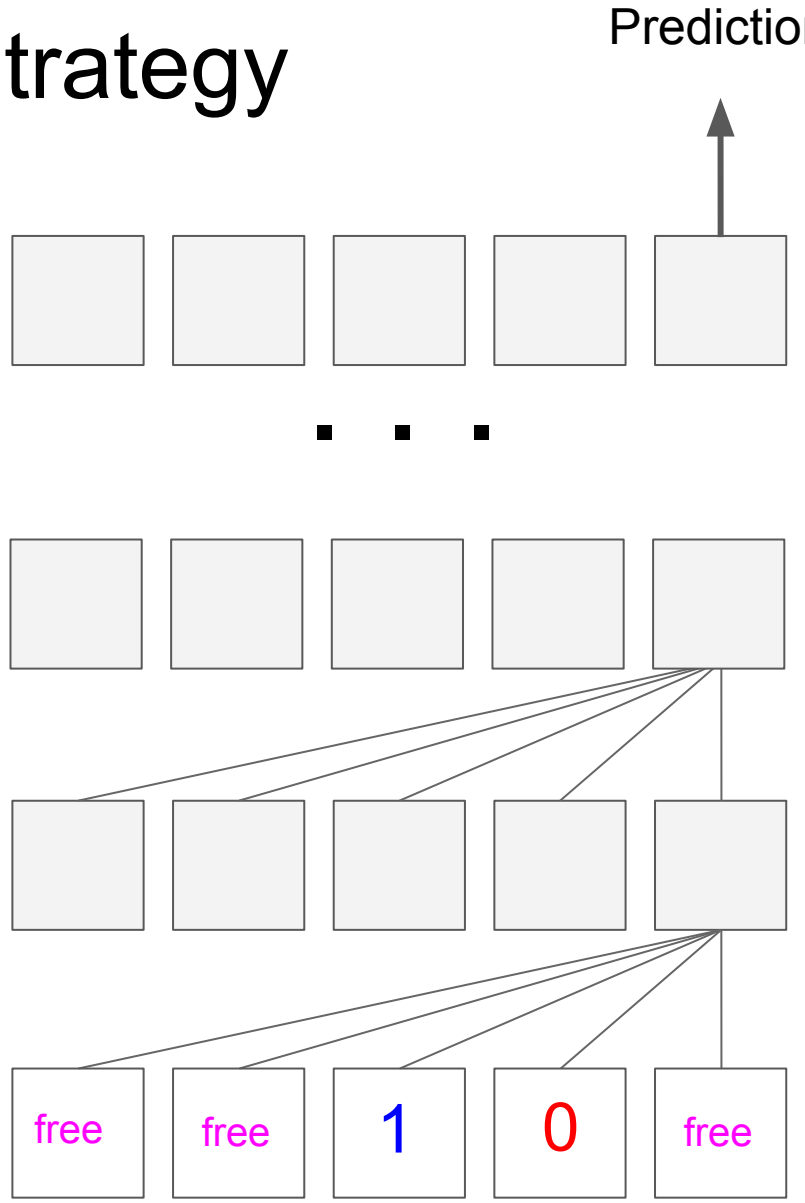
Enough to show that this is > 0 if the input string is sufficiently long!

Show this by

1. calculating for each head
2. combining via Lovasz Local Lemma

Hahn, TACL 2020

Proof Strategy



What is probability that the model ignores some free inputs?

Enough to show that this is > 0 if the input string is sufficiently long!

Show this by

1. calculating for each head
2. combining via Lovasz Local Lemma
3. induction over number of layers

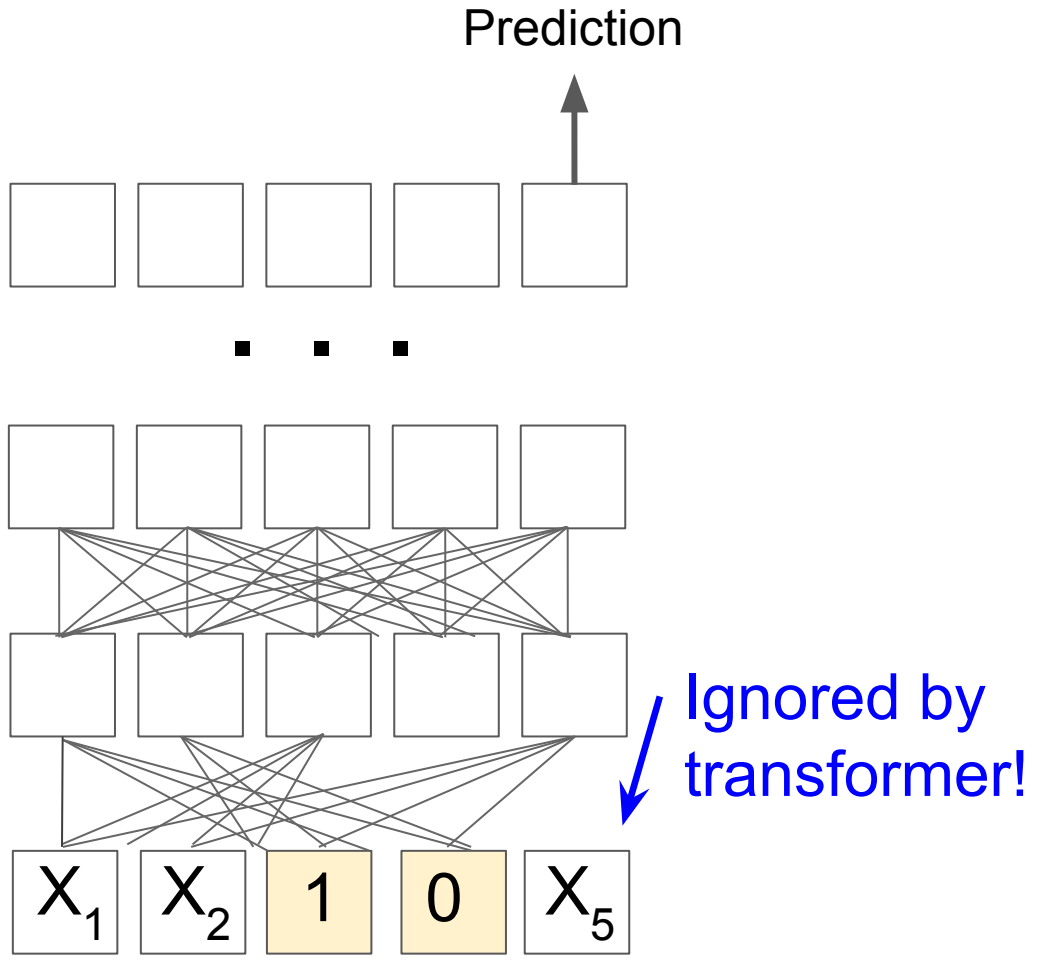
Hahn, TACL 2020

Proof Strategy

Assume we have a transformer for PARITY.

Construct two inputs that are classified the same, even though one is EVEN and the other is ODD.

Method: Strategically fix a few input bits to 'distract' the transformer, so that it ignores part of the input.



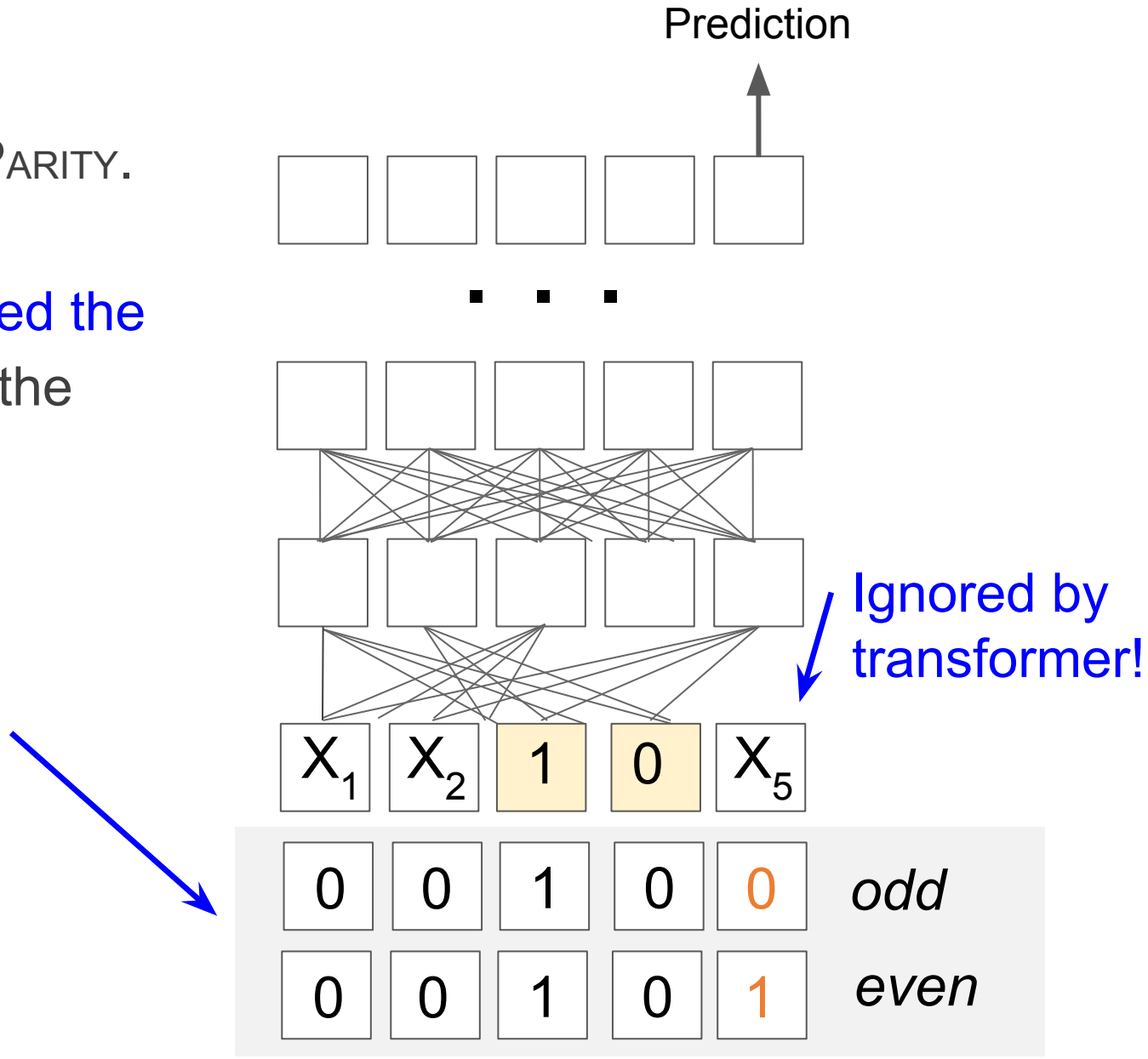
Hahn, TACL 2020

Proof Strategy

Assume we have a transformer for PARITY.

Construct two inputs that are classified the same, even though one is EVEN and the other is ODD.

These strings will be classified the same, even though their parities are different.



Hahn, TACL 2020

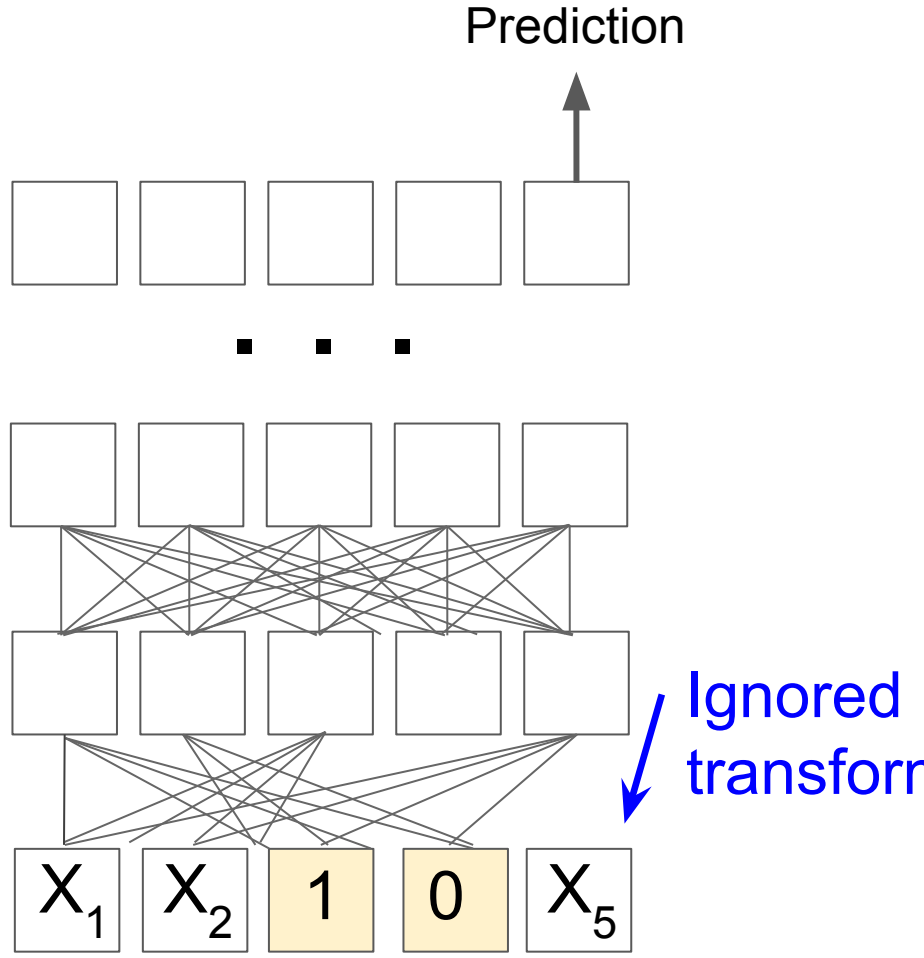
Proof Strategy

Assume we have a transformer for PARITY.

Construct two inputs that are classified the same, even though one is EVEN and the other is ODD.

These strings will be classified the same, even though their parities are different.

Contradiction!

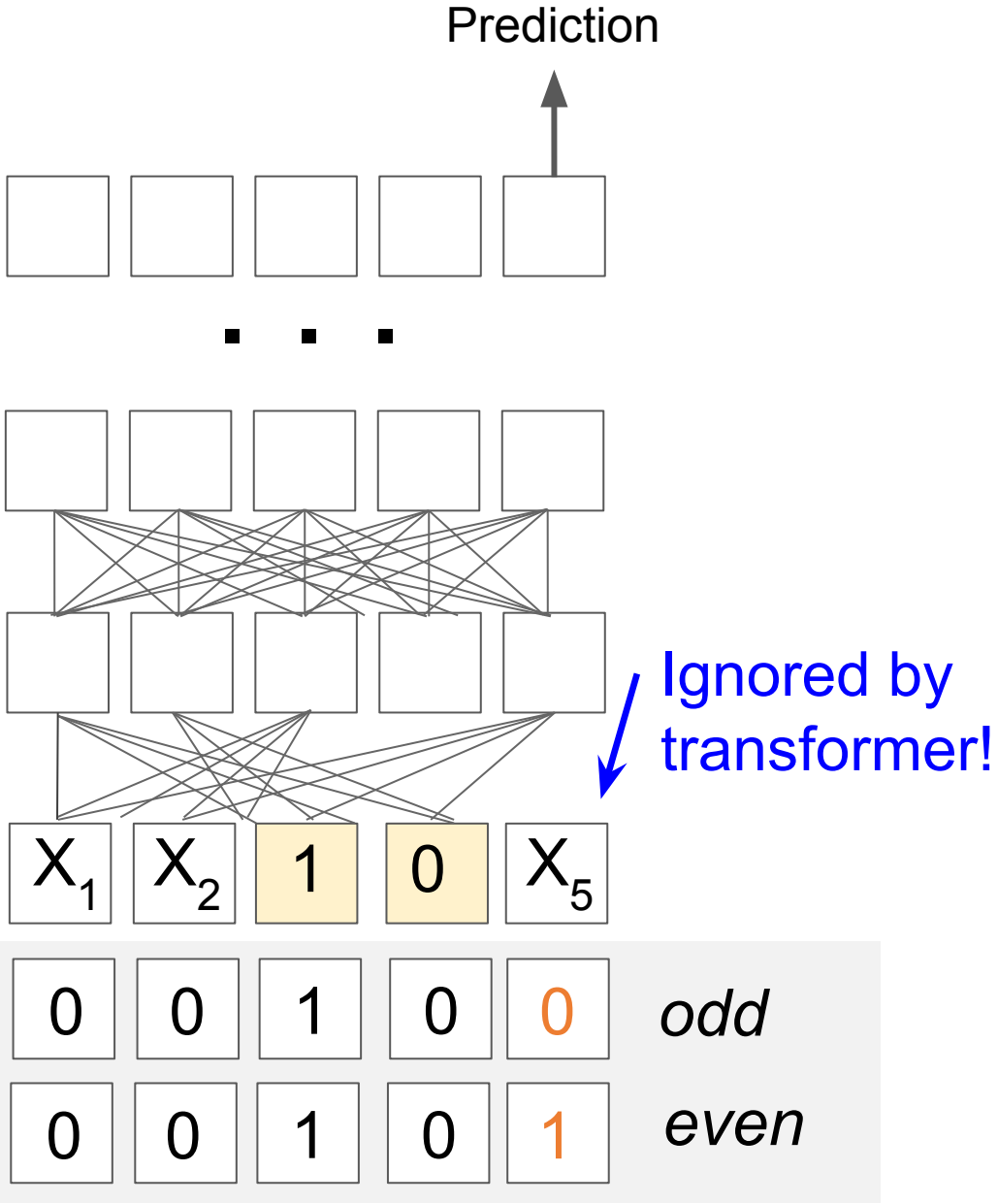


Hahn, TACL 2020

Limitations of Transformers

Theorem (TACL 2020)

No transformer can represent PARITY robustly at all input lengths.



Hahn, TACL 2020

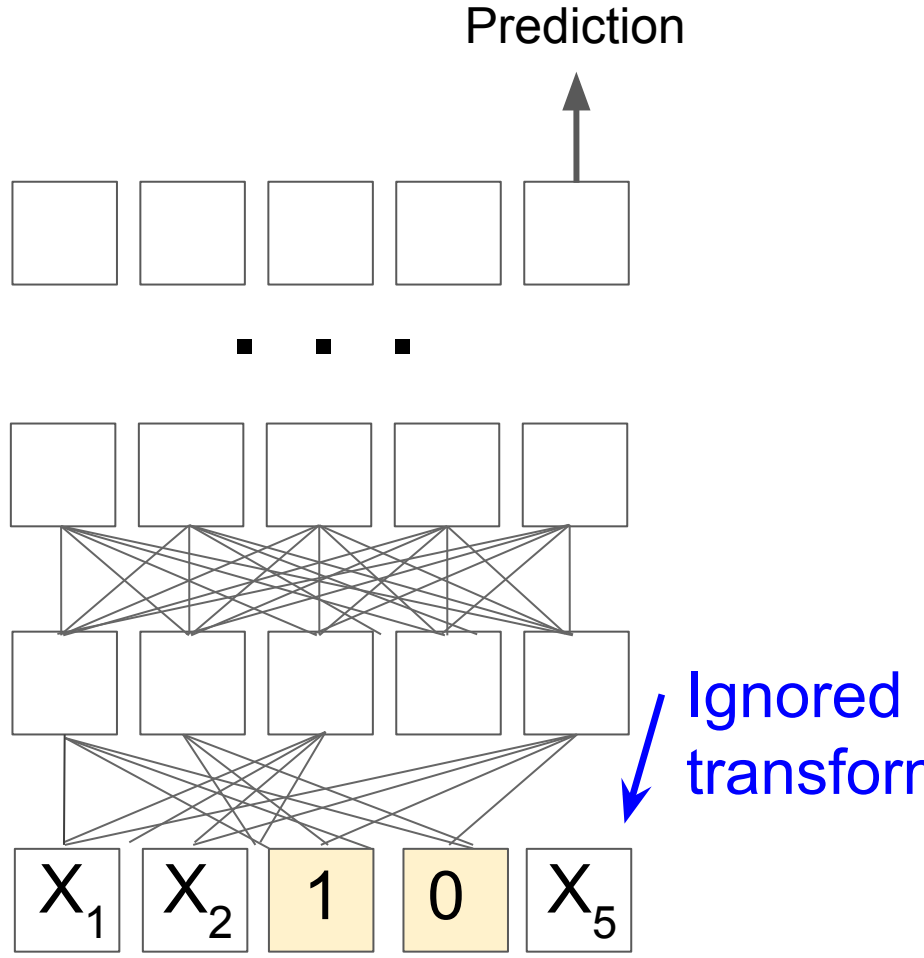
Limitations of Transformers

Theorem (TACL 2020)

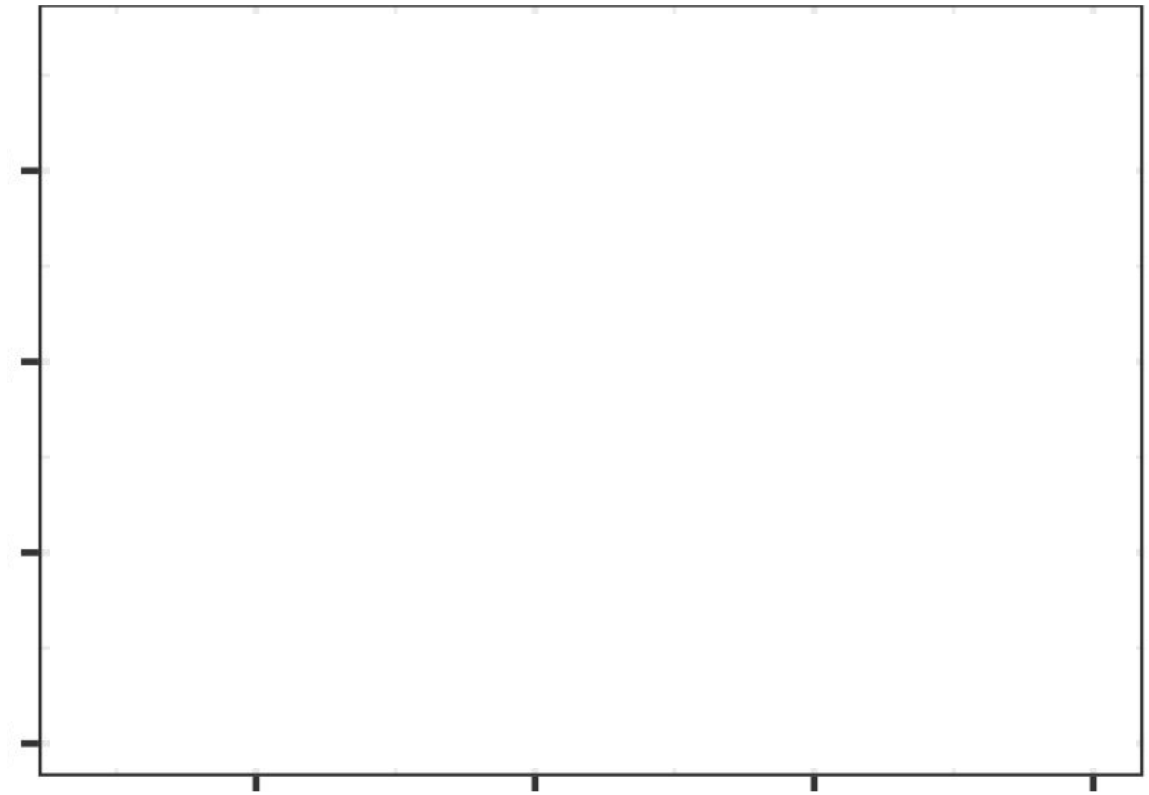
No transformer can represent PARITY robustly at all input lengths.

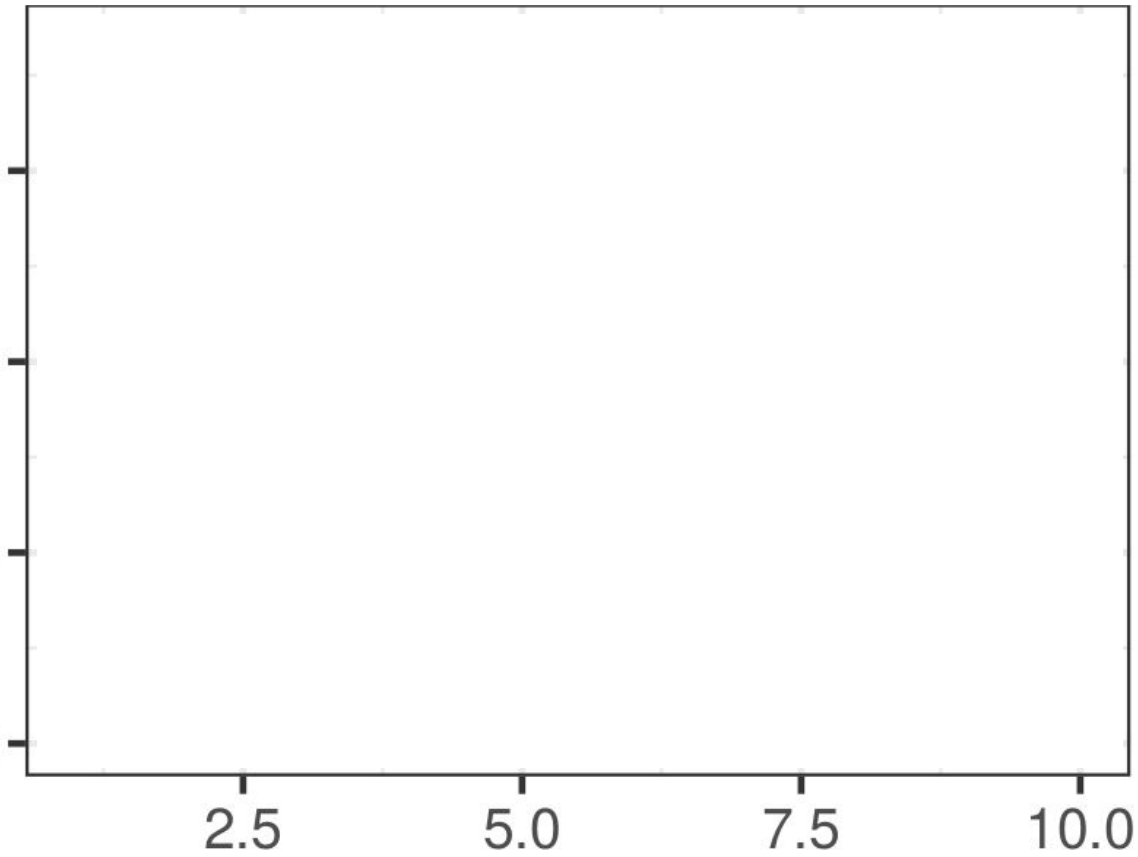
Consistently confirmed by empirical studies

- Bhattamishra et al 2020
- Chiang and Cholak 2021
- Delétang et al 2022
- Ruoss et al 2023



Hahn, TACL 2020

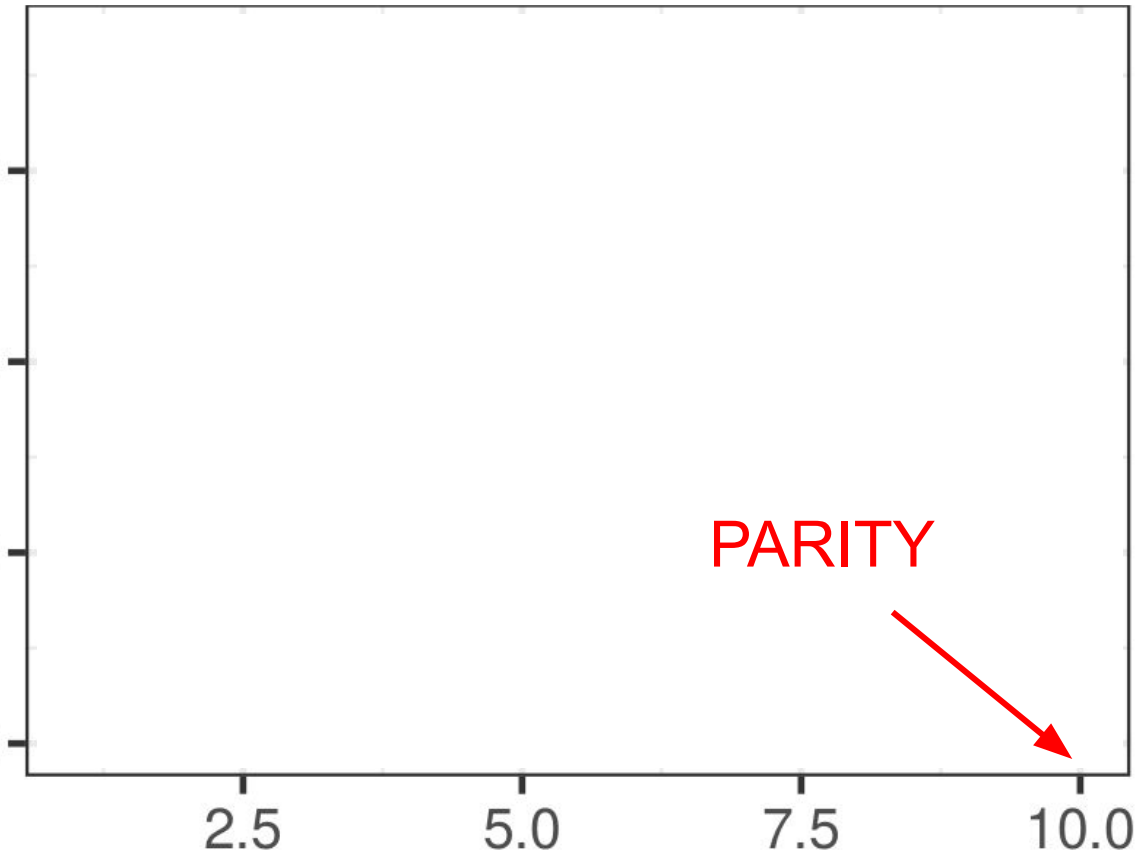




bitstrings of length 10

Average number of Hamming neighbors on which the label flips

("average sensitivity" or "total influence")

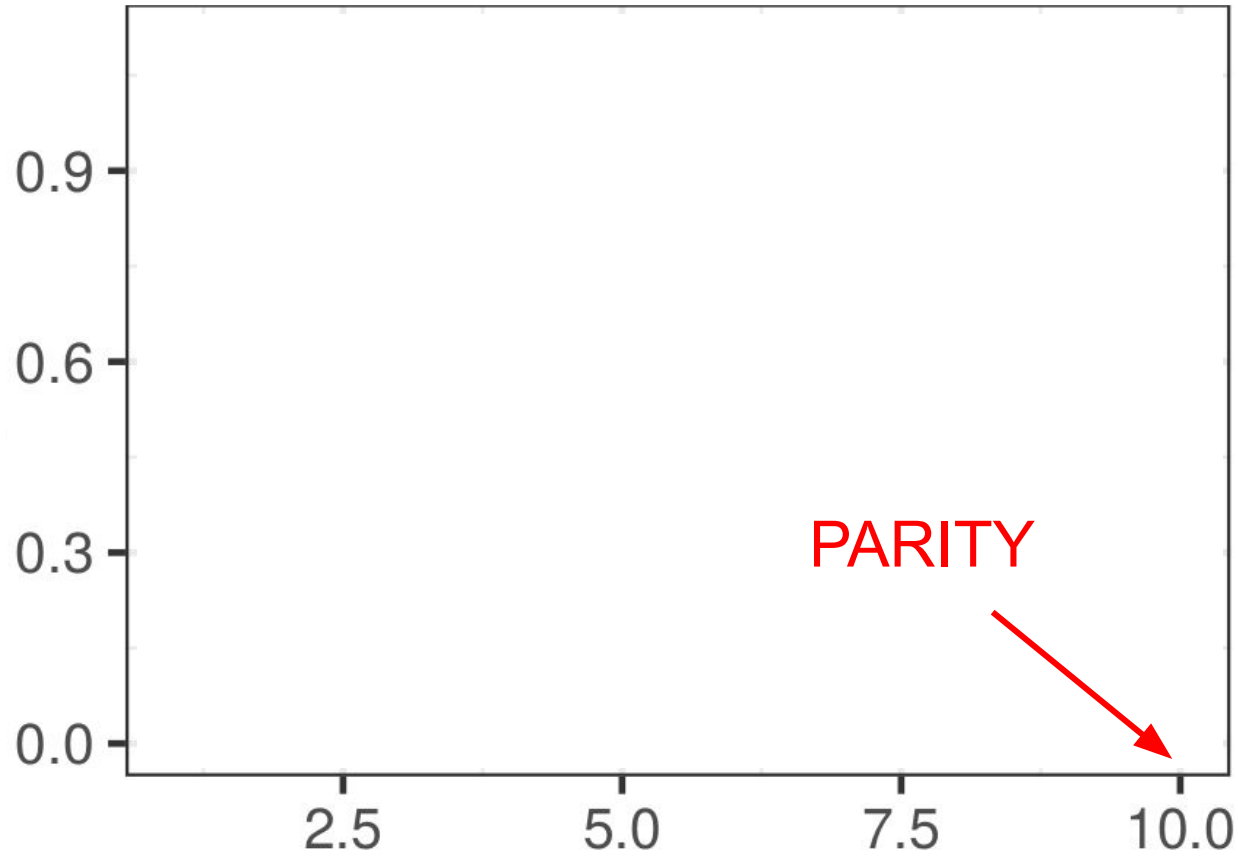


bitstrings of length 10

Average number of Hamming neighbors on which the label flips

("average sensitivity" or "total influence")

MSE on synthetic functions



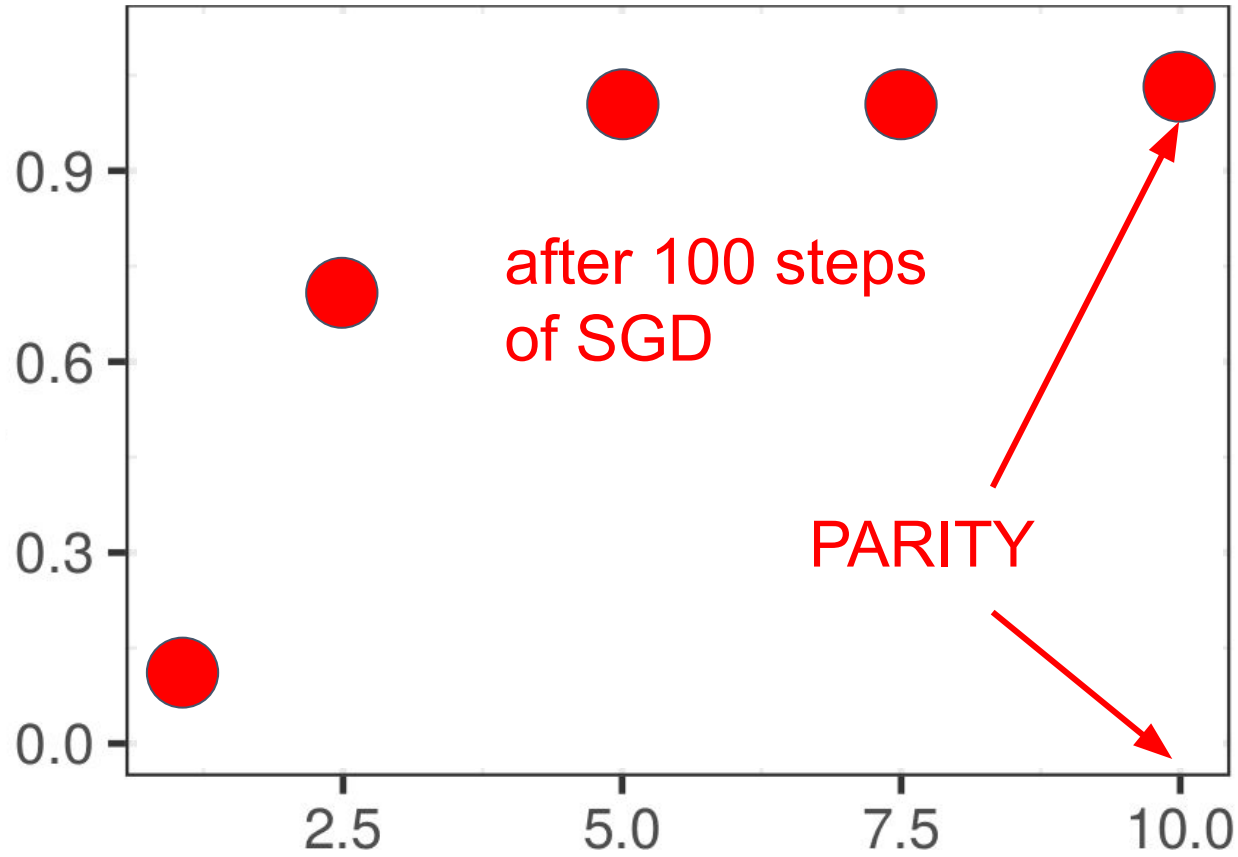
4 layers,
4 heads,
32 units
trained from scratch

bitstrings of length 10

Average number of Hamming neighbors on which the label flips

("average sensitivity" or "total influence")

MSE on synthetic functions



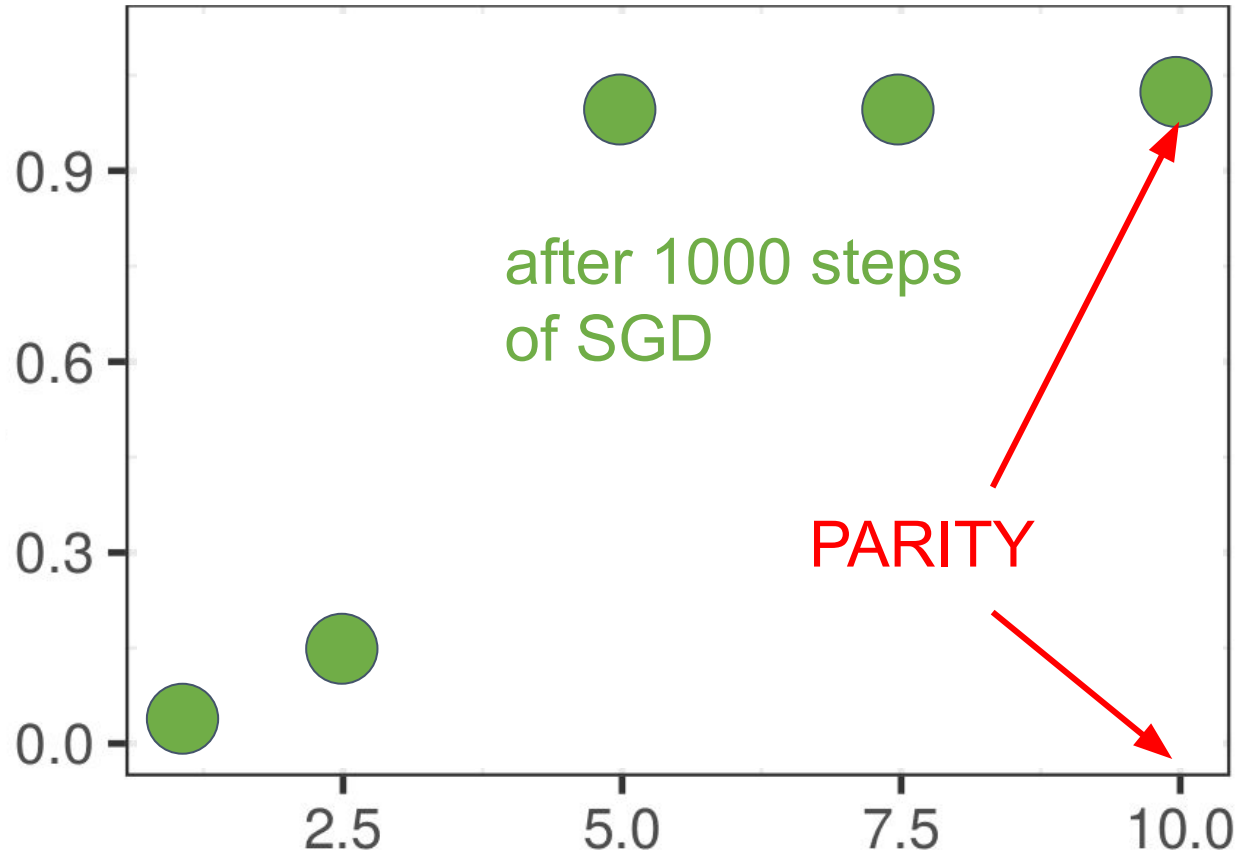
4 layers,
4 heads,
32 units
trained from scratch

bitstrings of length 10

Average number of Hamming neighbors on which the label flips

("average sensitivity" or "total influence")

MSE on synthetic functions



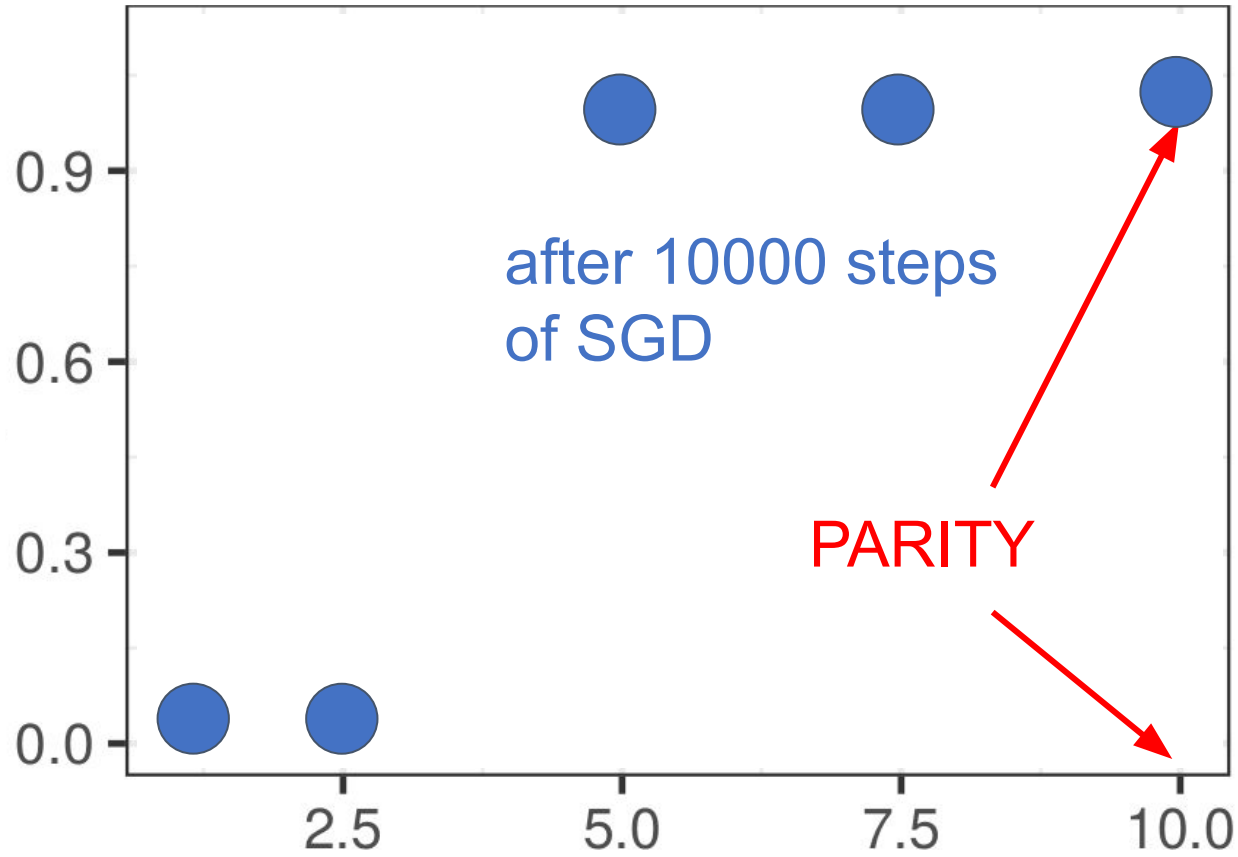
4 layers,
4 heads,
32 units
trained from scratch

bitstrings of length 10

Average number of Hamming neighbors on which the label flips

("average sensitivity" or "total influence")

MSE on synthetic functions



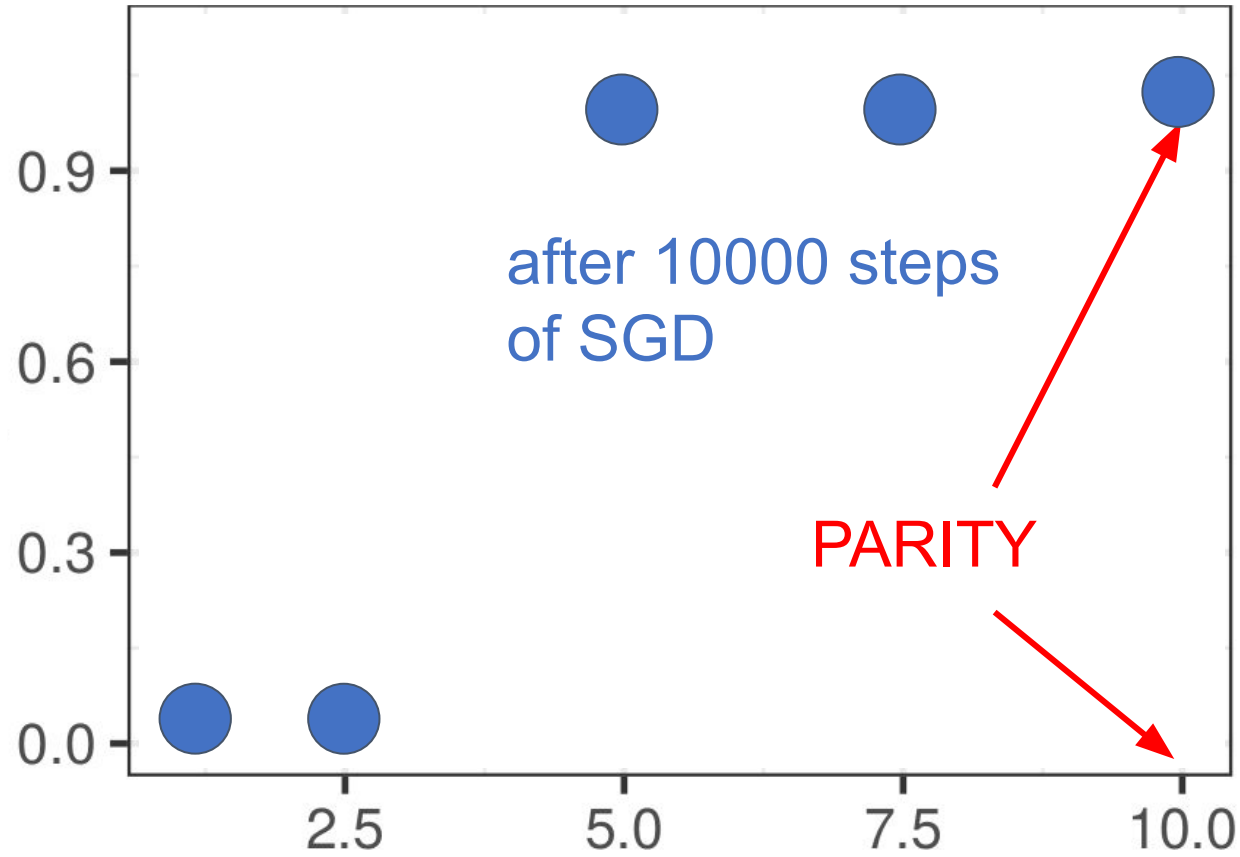
4 layers,
4 heads,
32 units
trained from scratch

bitstrings of length 10

Average number of Hamming neighbors on which the label flips

("average sensitivity" or "total influence")

MSE on synthetic functions

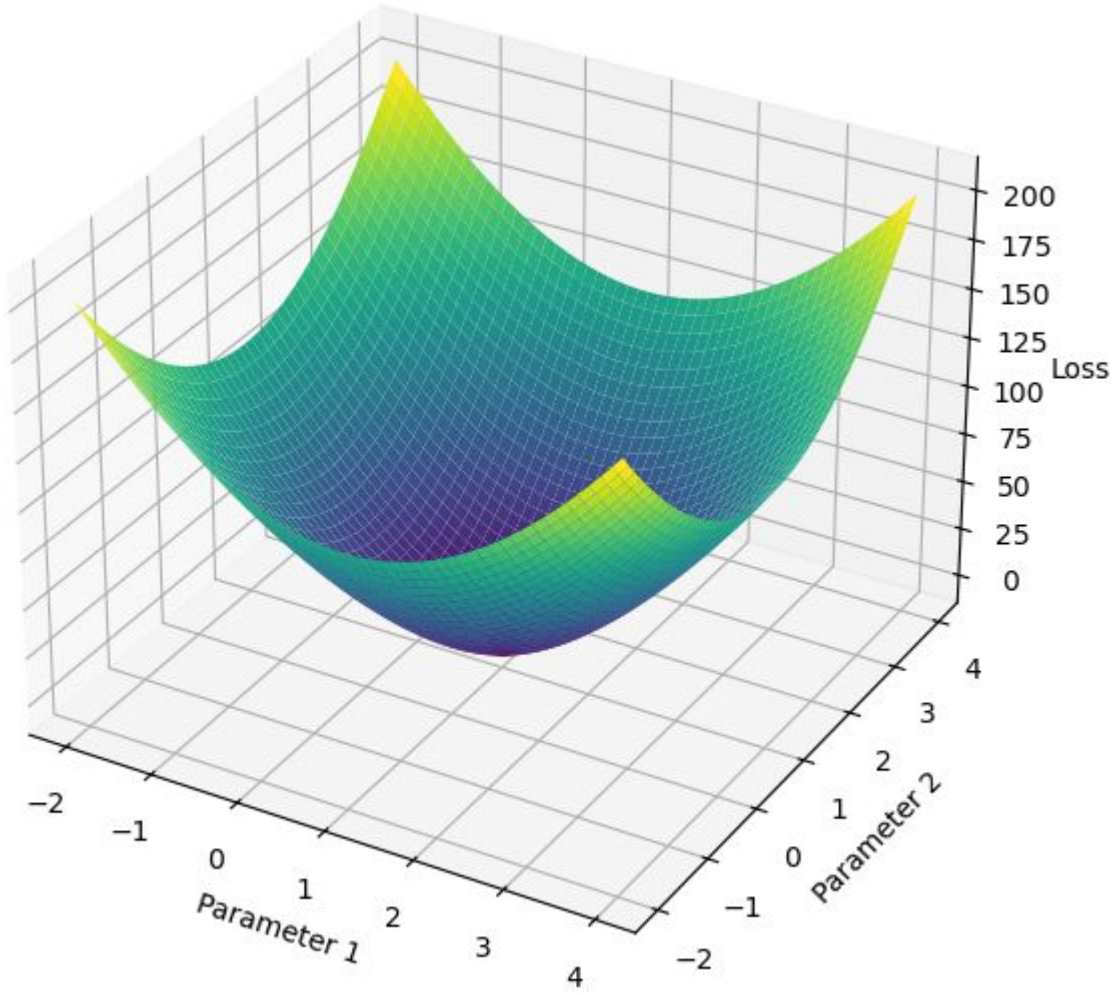


Highly sensitive functions are very hard for transformers to learn.

Average number of Hamming neighbors on which the label flips

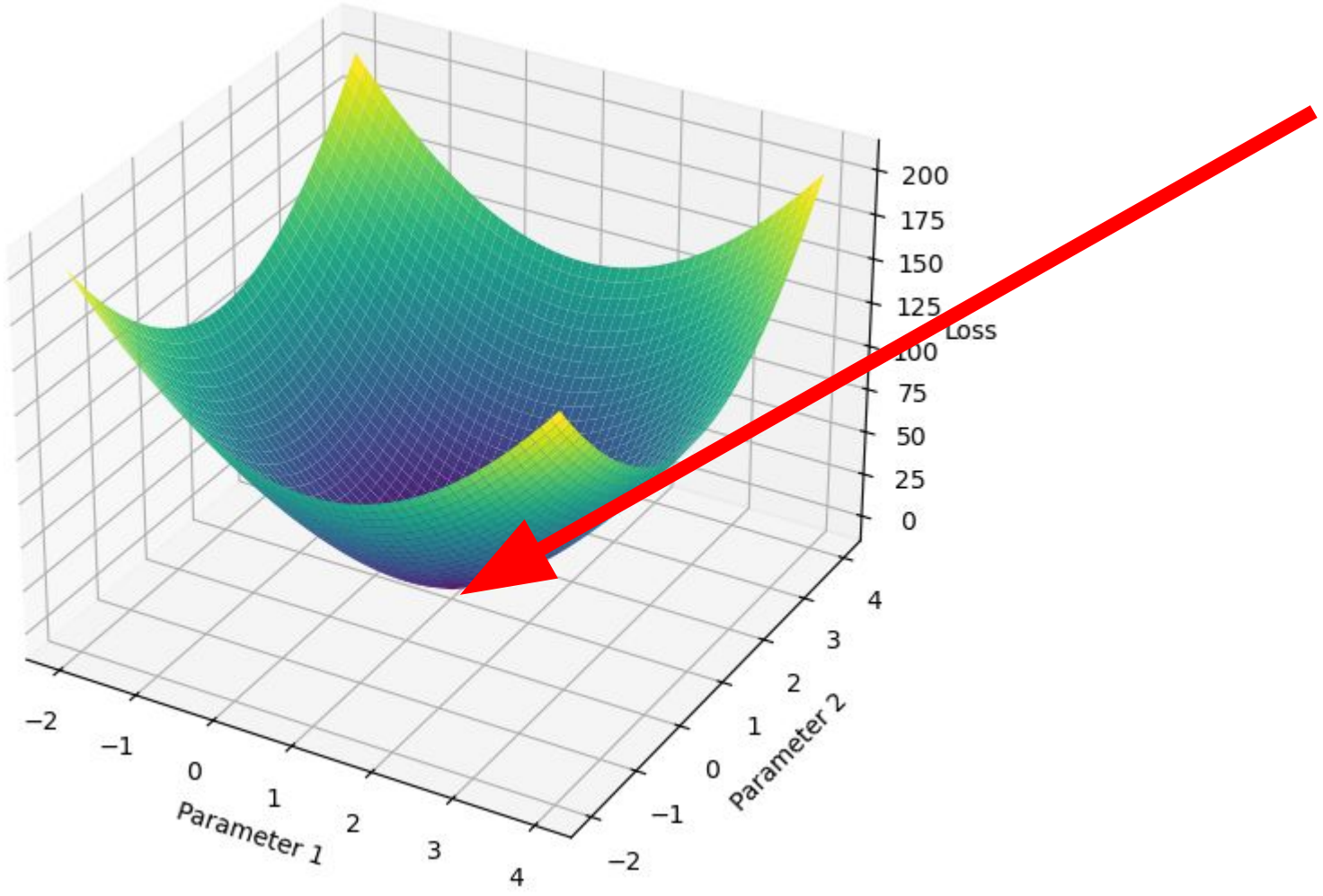
("average sensitivity" or "total influence")

Making this more Quantitative: Minima Sharpness



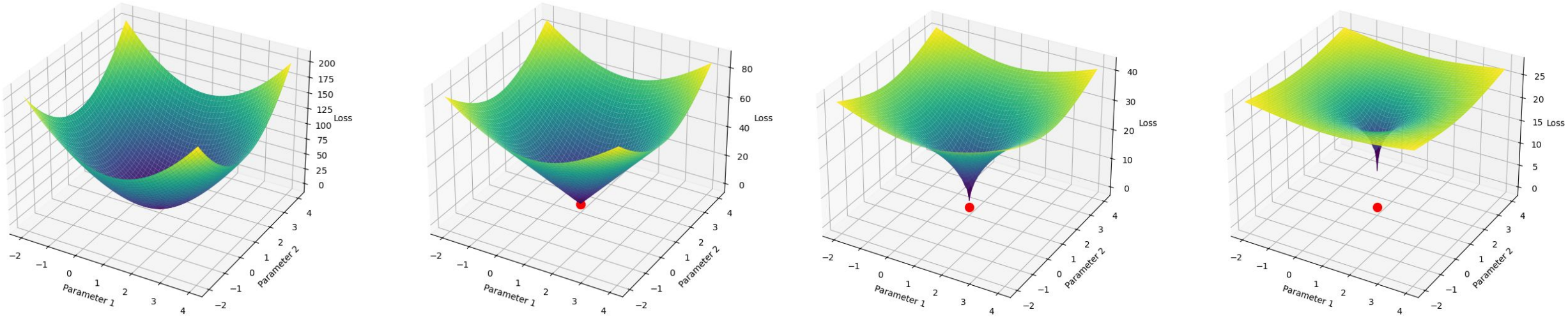
Hahn & Rofin, ACL 2024

Making this more Quantitative: Minima Sharpness



Hahn & Rofin, ACL 2024

Making this more Quantitative: Minima Sharpness

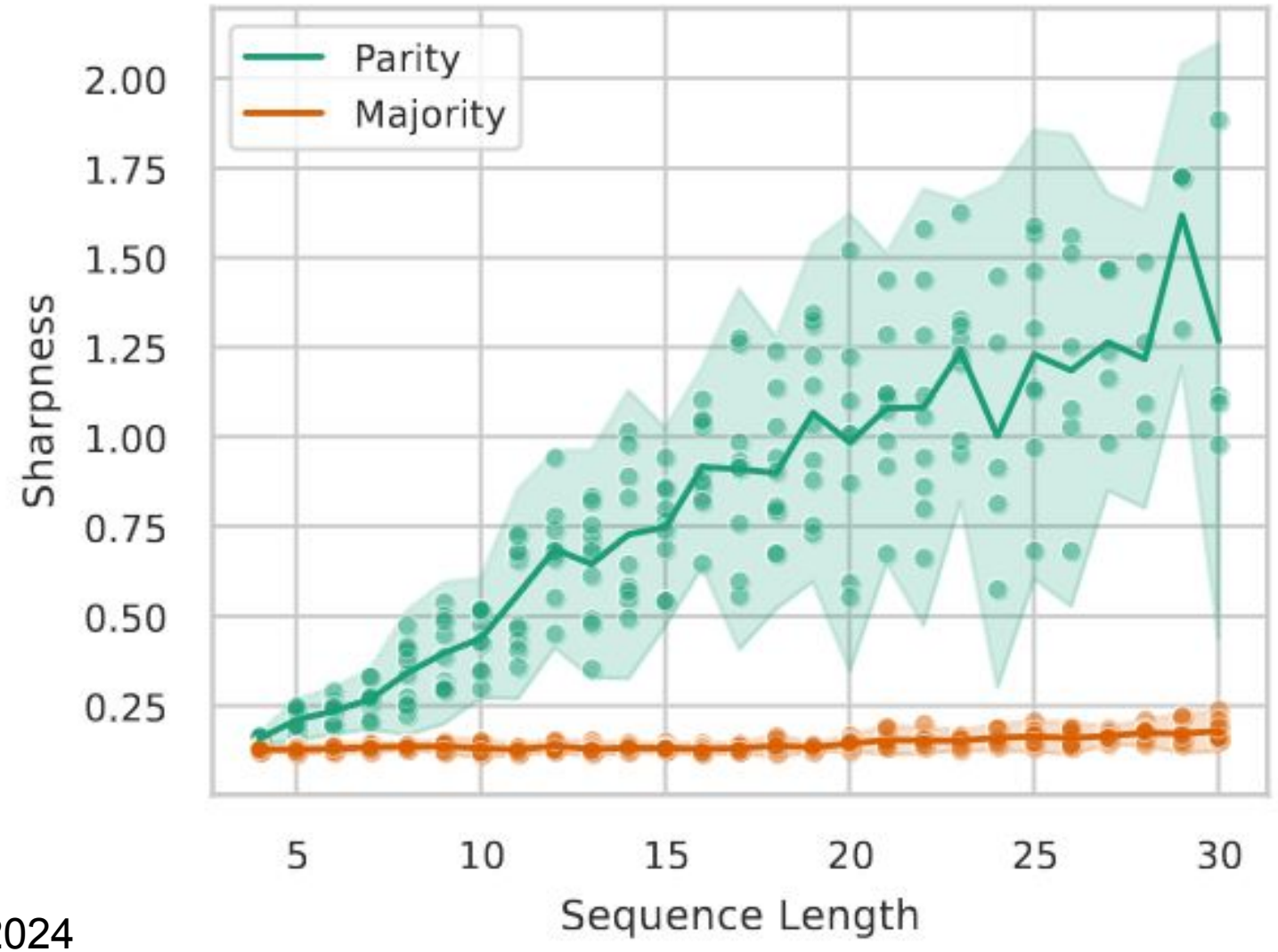


more sensitive functions

sharper minima

Hahn & Rofin, ACL 2024

Making this more Quantitative: Minima Sharpness



Hahn & Rofin, ACL 2024

M **You**

What is the parity of the following bit string? 0111011010

0111011010

not cherry-picked!
Just typed a random bit string of length 10.

0111011010

six 1’s --> even parity

not cherry-picked!
Just typed a random bit string of length 10.

You

What is the parity of the following bit string? 0111011010

ChatGPT

To determine the parity of a bit string, you count the number of 1s in the string. If the count of 1s is even, the parity is even, and if the count is odd, the parity is odd.

You

What is the parity of the following bit string? 0111011010

ChatGPT

To determine the parity of a bit string, you count the number of 1s in the string. If the count of 1s is even, the parity is even, and if the count is odd, the parity is odd.

In the bit string "0111011010," there are 5 ones. Since 5 is an odd number, the parity of the bit string is odd.

You

What is the parity of the following bit string? 0111011010

ChatGPT

To determine the parity of a bit string, you count the number of 1s in the string. If the count of 1s is even, the parity is even, and if the count is odd, the parity is odd.

In the bit string "0111011010," there are **5** ones. Since 5 is an odd number, the parity of the bit string is odd.

!

It is even inconsistent across runs

You

What is the parity of the following bit string? 0111011010

ChatGPT

To determine the parity of a bit string, you need to count the number of '1' bits in the string. If the count is even, the parity is even; if the count is odd, the parity is odd.

Let's count the '1' bits in the given bit string: 0111011010

There are five '1' bits in the string. Since 5 is an odd number, the parity of the bit string is odd.

5 odd

It is even inconsistent across runs

You
What is the parity of the following bit string? 0111011010

ChatGPT
To determine the parity of a bit string, you need to count the number of '1' bits in the string. If the count is even, the parity is even; if the count is odd, the parity is odd.

Let's count the '1' bits in the given bit string: 0111011010

There are five '1' bits in the string. Since 5 is an odd number, the parity of the bit string is odd.

5 odd

You
What is the parity of the following bit string? 0111011010

ChatGPT
To determine the parity of a bit string, you count the number of 1s in the string. If the count is even, the parity is even; if the count is odd, the parity is odd.

Let's count the number of 1s in the given bit string: 0111011010.

1. Count of 1s: 6

Since the count of 1s is even (6 is an even number), the parity of the given bit string is even.

6 even

It is even inconsistent across runs

You
What is the parity of the following bit string? 0111011010

ChatGPT
To determine the parity of a bit string, you need to count the number of '1' bits in the string. If the count is even, the parity is even; if the count is odd, the parity is odd.

Let's count the '1' bits in the given bit string: 0111011010

There are five '1' bits in the string. Since 5 is an odd number, the parity of the bit string is odd.

5 odd

You
What is the parity of the following bit string? 0111011010

ChatGPT
To determine the parity of a bit string, you count the number of ones in the string. If the count is even, the parity is even; if the count is odd, the parity is odd.

Let's count the number of ones in the given bit string: 0111011010.

Number of ones: 1 + 1 + 1 + 1 + 1 + 1 + 0 + 1 + 0 + 0 = 7

Since the count of ones is odd (7), the parity of the bit string is odd.

7 odd

You
What is the parity of the following bit string? 0111011010

ChatGPT
To determine the parity of a bit string, you count the number of 1s in the string. If the count is even, the parity is even; if the count is odd, the parity is odd.

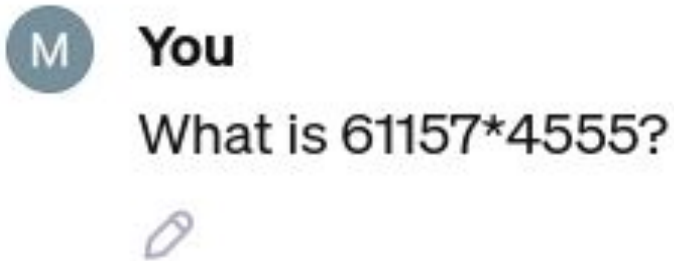
Let's count the number of 1s in the given bit string: 0111011010.

1. Count of 1s: 6

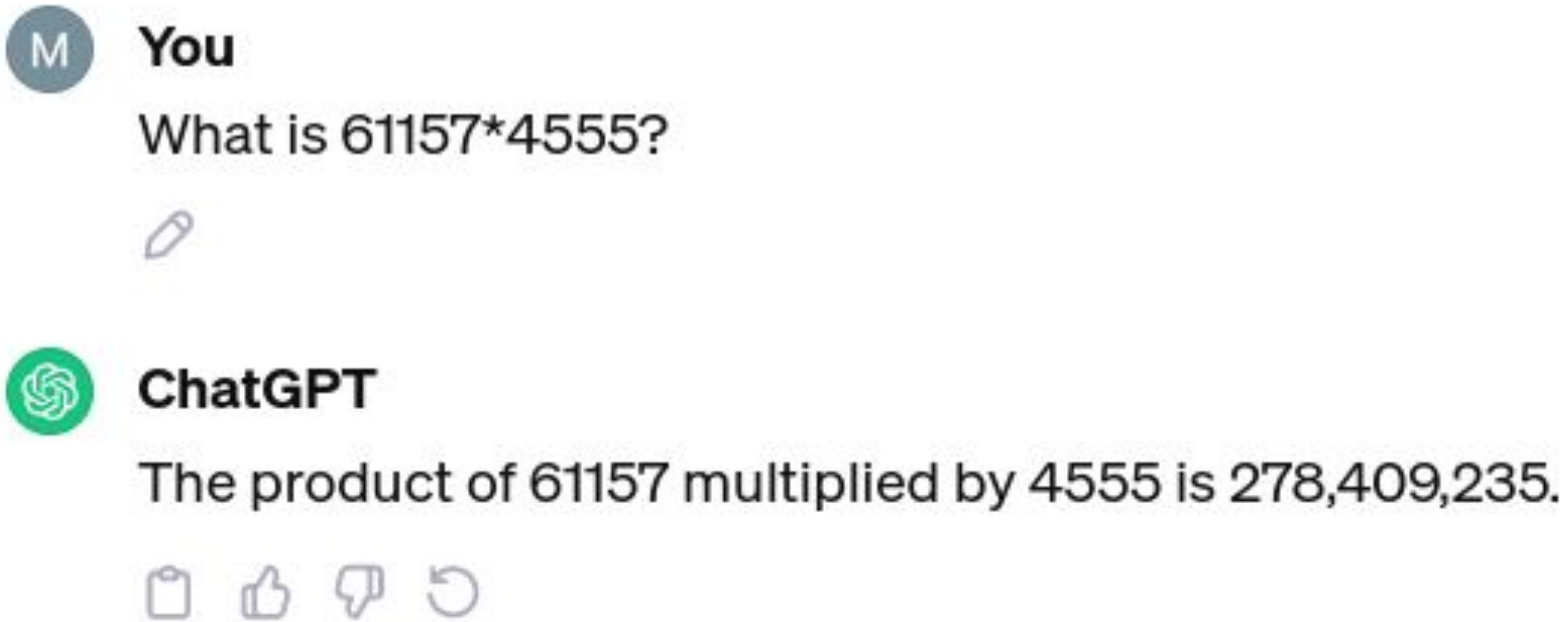
Since the count of 1s is even (6 is an even number), the parity of the given bit string is even.

6 even

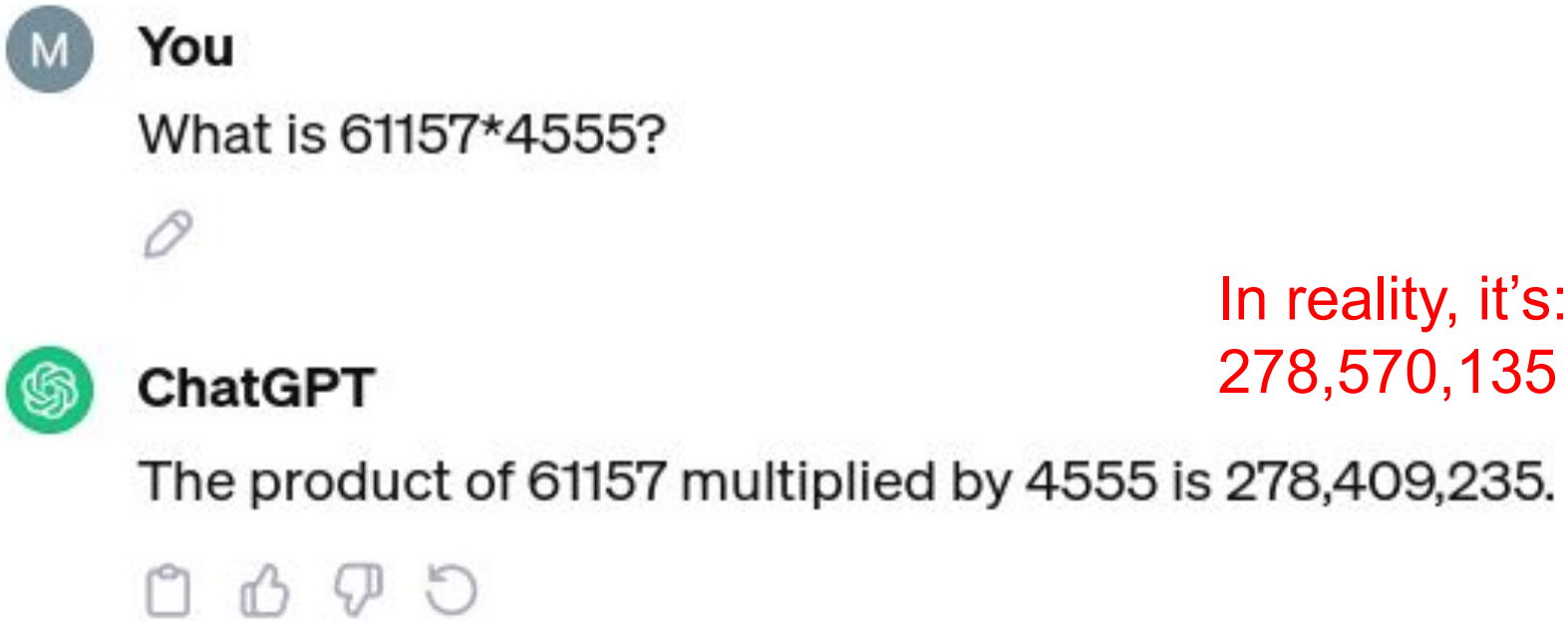
Not just parity



Not just parity



Not just parity



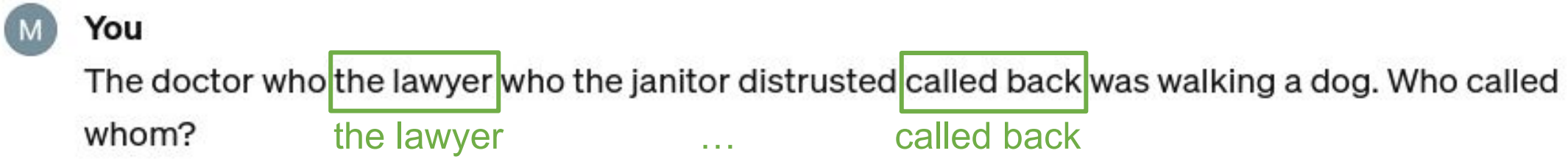
In reality, it's:
278,570,135

Not just parity

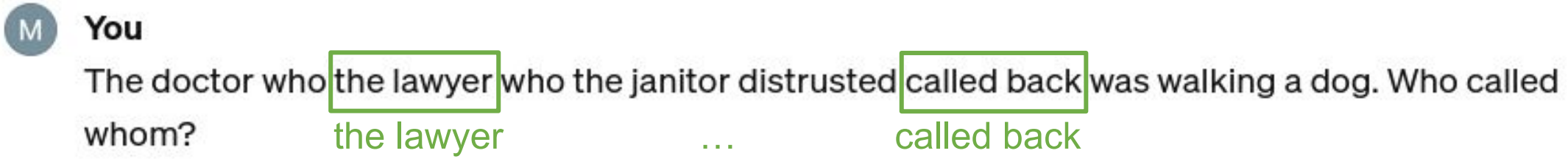
M **You**

The doctor who the lawyer who the janitor distrusted called back was walking a dog. Who called whom?

Not just parity



Not just parity



Getting this right requires context-free parsing.
Similar limitations apply as for PARITY Hahn, TACL 2020.

Not just parity

You

The doctor who the lawyer who the janitor distrusted called back was walking a dog. Who called whom?

the lawyer … called back

ChatGPT

The sentence is a bit complex, but let's break it down:

1. The janitor distrusted the lawyer.
2. The lawyer distrusted the doctor.
3. The doctor who the lawyer distrusted called back.

So, in the end, it was the doctor who called back. The person walking the dog is the doctor whom the lawyer distrusted.

Not just parity

You

The doctor who the lawyer who the janitor distrusted called back was walking a dog. Who called whom?

the lawyer … called back

ChatGPT

The sentence is a bit complex, but let's break it down:

1. The janitor distrusted the lawyer.
2. The lawyer distrusted the doctor.
3. The doctor who the lawyer distrusted called back.

So, in the end, it was the doctor who called back. The person walking the dog is the doctor whom the lawyer distrusted.

Not just parity

You

The doctor who the lawyer who the janitor distrusted called back was walking a dog. Who called whom?

the lawyer … called back

ChatGPT

The sentence is a bit complex, but let's break it down:

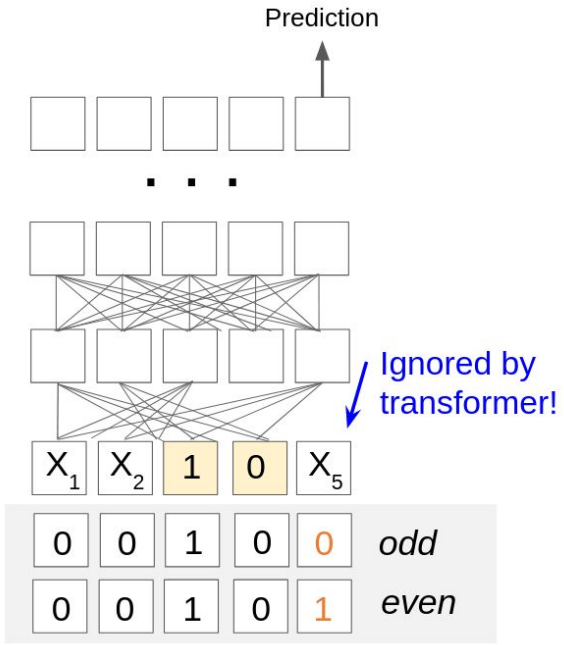
1. The janitor distrusted the lawyer.
2. The lawyer distrusted the doctor.
3. The doctor who the lawyer distrusted called back.

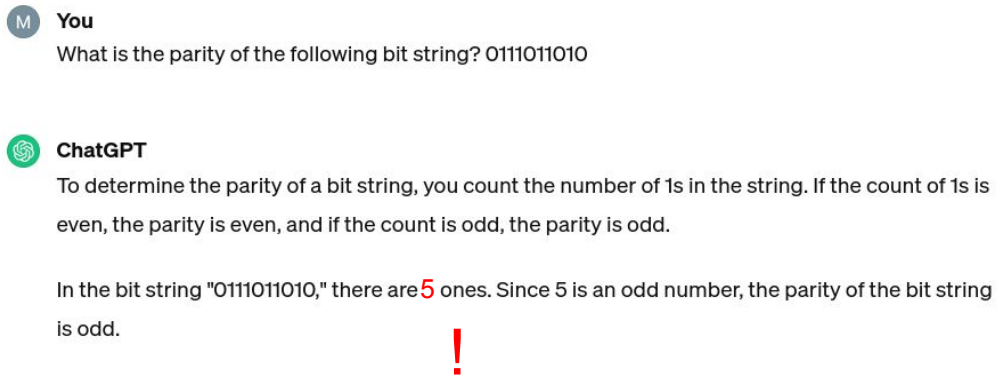
incorrect!

So, in the end, it was the doctor who called back. The person walking the dog is the doctor whom the lawyer distrusted.

Take-Away

Theory uncovers limitations that are relevant even for today's LLMs.

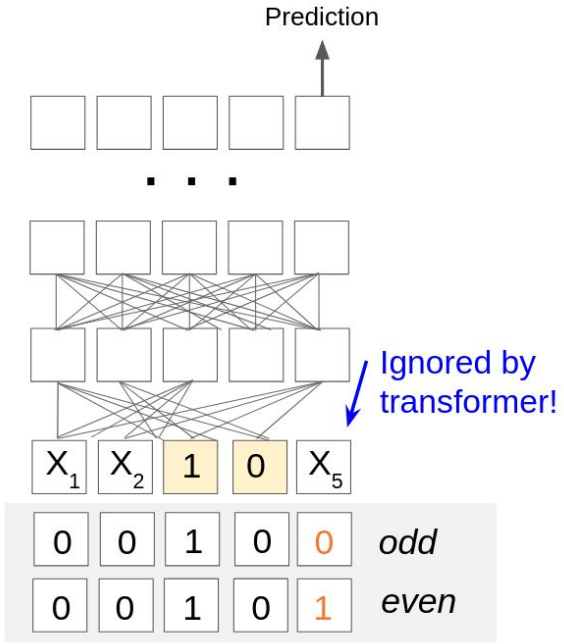


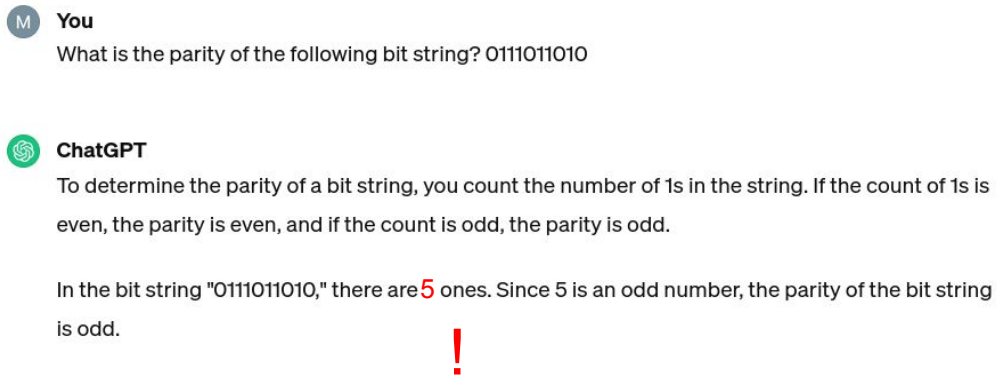


Take-Away

Theory uncovers limitations that are relevant even for today's LLMs.

Transformers struggle with highly sensitive computations.





What are the computational mechanisms needed for language understanding?

What are the in-principle capabilities and limitations of neural ML architectures?

HG, in subm. 2023

HJF, TACL 2021

H**H**GLM, EMNLP 2021

HB, TACL 2021

H, TACL 2020

Can we use LMs to reverse-engineer human language comprehension?

HK, Cognition 2023

HFLG, PNAS 2022

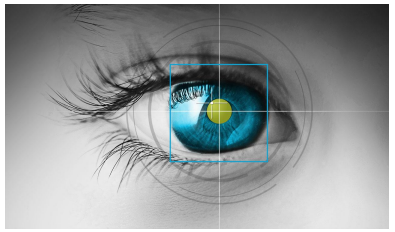
HJF, PNAS 2020

HFBB, CogSci 2019

HK, EMNLP 2016

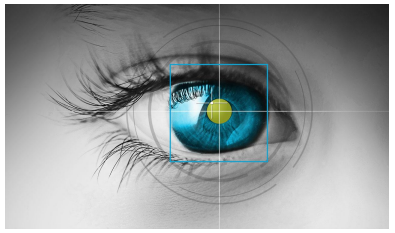
Human Reading

Human Reading



The two young sea-lions took not the slightest interest in our arrival. They were playing on the jetty, rolling over and tumbling into the water together, entirely

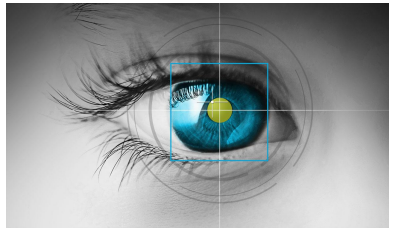
Human Reading



The two young sea-lions took not the slightest interest in our arrival. They were playing on the jetty, rolling over and tumbling into the water together, entirely

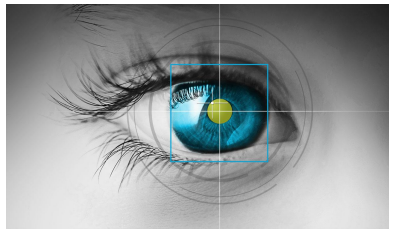
Human Reading

saccade



The two young sea-lions took not the slightest interest in our arrival. They were playing on the jetty, rolling over and tumbling into the water together, entirely

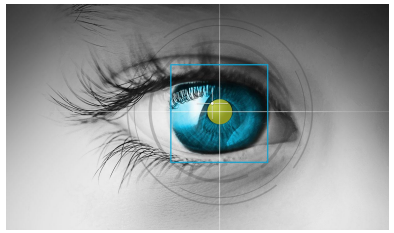
Human Reading



fixation

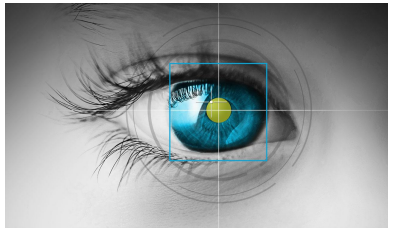
The two young sea-lions took not the slightest interest in our arrival. They were playing on the jetty, rolling over and tumbling into the water together, entirely

Human Reading



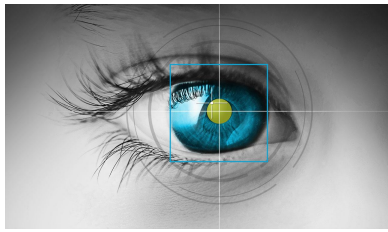
The two young sea-lions took not the slightest interest in our arrival. They were playing on the jetty, rolling over and tumbling into the water together, entirely

Human Reading



The two young sea-lions took not the slightest interest in our arrival. They were playing on the jetty, rolling over and tumbling into the water together, entirely

Human Reading



The two young sea-lions took not the slightest interest in our arrival. They were playing on the jetty, rolling over and tumbling into the water together, entirely

Human Reading

The two young sea-lions took not the slightest interest in our arrival. They were playing on the jetty, rolling over and tumbling into the water together, entirely

Modeling Human Reading

Interest in human reading for NLP

(e.g., Barrett et al 2018; Malmaud et al 2020; Sood et al 2020)

Modeling Human Reading

Interest in human reading for NLP (e.g., Barrett et al 2018; Malmaud et al 2020; Sood et al 2020)

Prior approaches:

supervised models (Nilsson and Nivre 2009; Mathies and Sogaard 2013)

Modeling Human Reading

Interest in human reading for NLP (e.g., Barrett et al 2018; Malmaud et al 2020; Sood et al 2020)

Prior approaches:

supervised models (Nilsson and Nivre 2009; Mathies and Sogaard 2013)

hand-crafted models (Engbert et al 2002; Reichle et al 1998, 2009)

Modeling Human Reading

Interest in human reading for NLP (e.g., Barrett et al 2018; Malmaud et al 2020; Sood et al 2020)

Prior approaches:

- supervised models (Nilsson and Nivre 2009; Mathies and Sogaard 2013)
- hand-crafted models (Engbert et al 2002; Reichle et al 1998, 2009)

Limitations:

- Need human gaze data

Modeling Human Reading

Interest in human reading for NLP (e.g., Barrett et al 2018; Malmaud et al 2020; Sood et al 2020)

Prior approaches:

- supervised models (Nilsson and Nivre 2009; Mathies and Sogaard 2013)
- hand-crafted models (Engbert et al 2002; Reichle et al 1998, 2009)

Limitations:

- Need human gaze data
- Does not generalize across reading tasks

Modeling Human Reading

Interest in human reading for NLP (e.g., Barrett et al 2018; Malmaud et al 2020; Sood et al 2020)

Prior approaches:

- supervised models (Nilsson and Nivre 2009; Mathies and Sogaard 2013)
- hand-crafted models (Engbert et al 2002; Reichle et al 1998, 2009)

Limitations:

- Need human gaze data
- Does not generalize across reading tasks

My approach:

- Unsupervised model overcoming limitations

Modeling Human Reading

Modeling Human Reading

Hypothesis: Human reading is optimized for

Modeling Human Reading

Hypothesis: Human reading is optimized for

efficiently extracting
information

Modeling Human Reading

Hypothesis: Human reading is optimized for

efficiently extracting information

while

maximizing speed

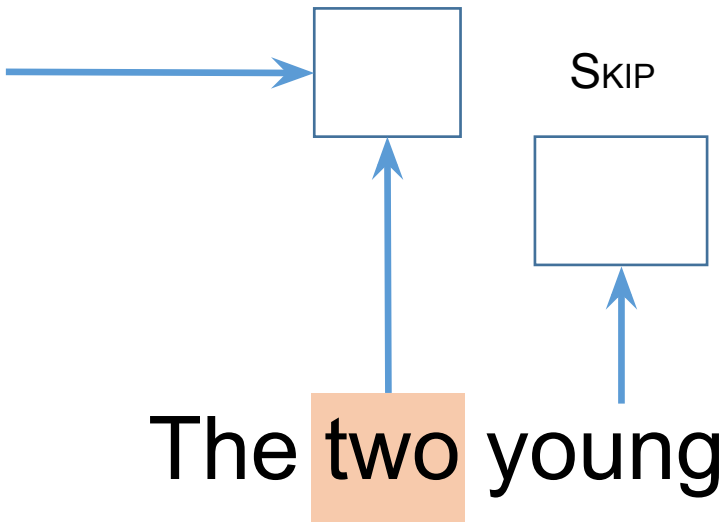
NEural Attention Tradeoff Model (NEAT)

The two young sea-lions took not the slightest interest

NEural Attention Tradeoff Model (NEAT)

Reader Network
reads text

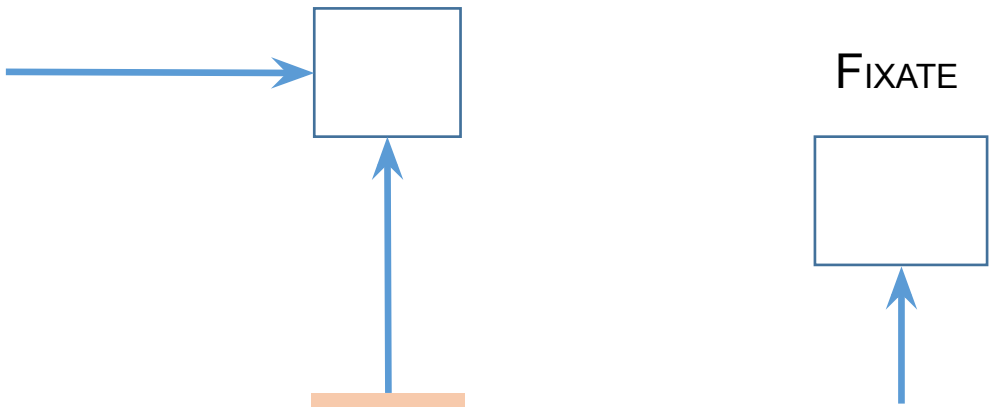
Control Network
decides which word to read or skip



The two young sea-lions took not the slightest interest

NEural Attention Tradeoff Model (NEAT)

Reader Network
reads text



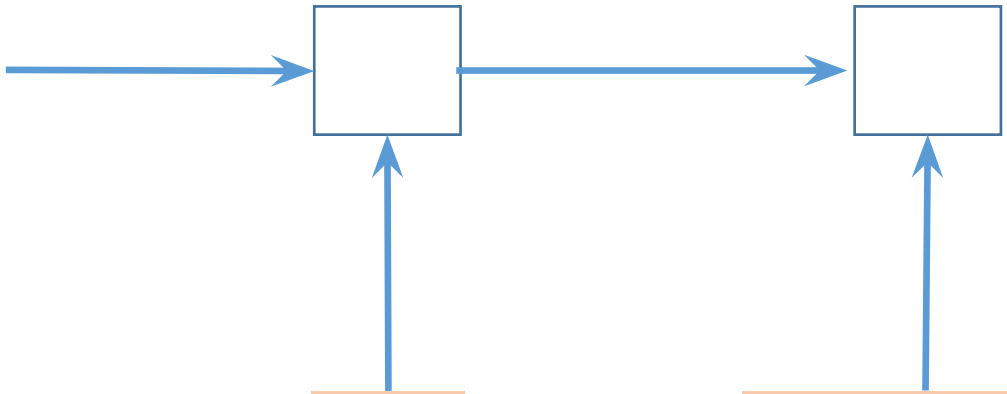
FIXATE

Control Network
decides which word to read or skip

The two young sea-lions took not the slightest interest

NEural Attention Tradeoff Model (NEAT)

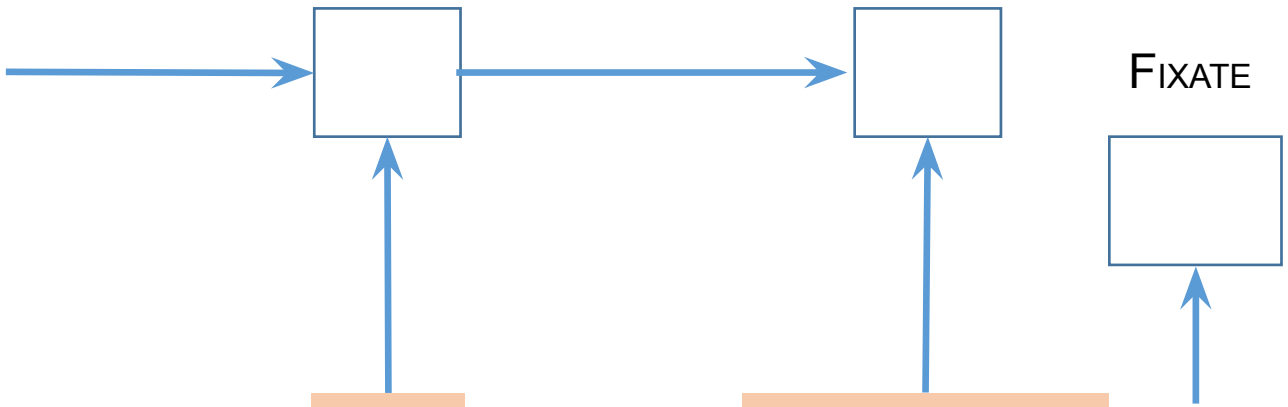
Reader Network
reads text



The two young sea-lions took not the slightest interest

NEural Attention Tradeoff Model (NEAT)

Reader Network
reads text

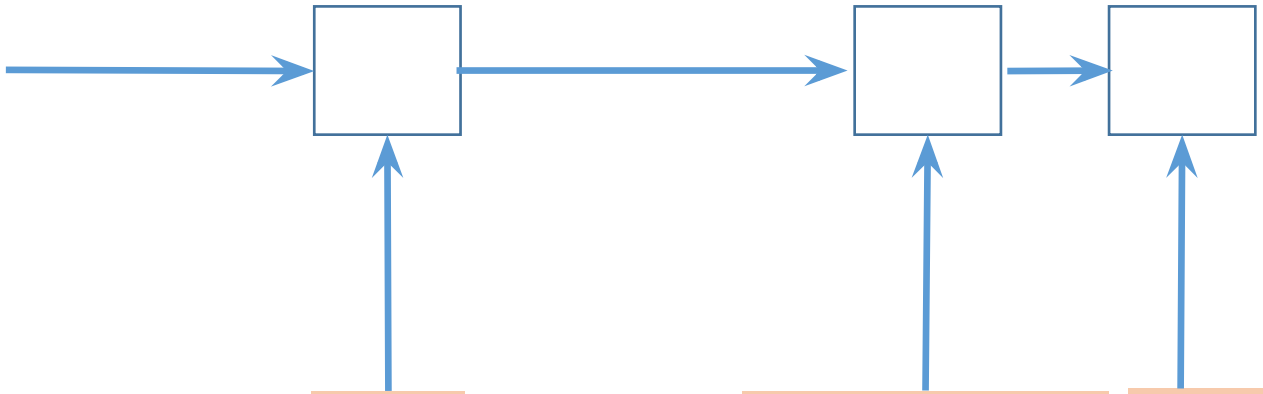


Control Network
decides which word to read or skip

The two young sea-lions took not the slightest interest

NEural Attention Tradeoff Model (NEAT)

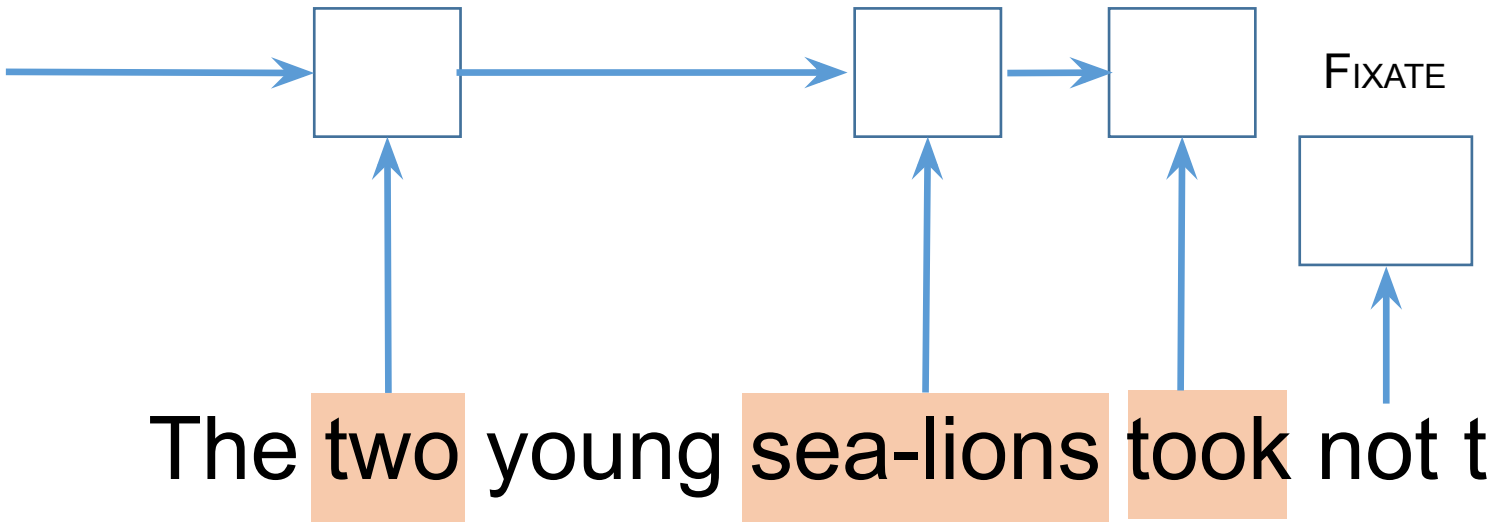
Reader Network
reads text



The two young sea-lions took not the slightest interest

NEural Attention Tradeoff Model (NEAT)

Reader Network
reads text

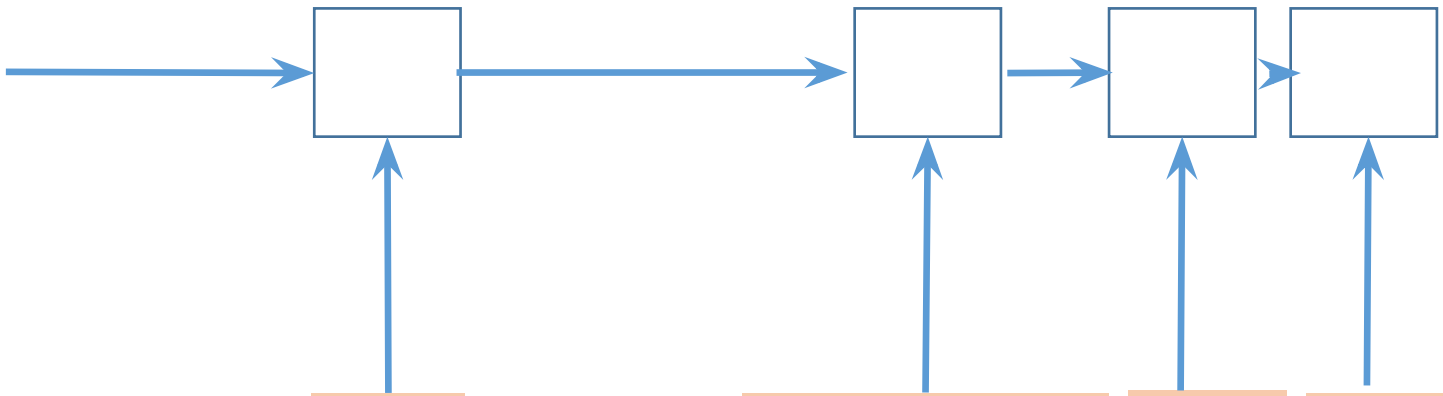


Control Network
decides which word to read or skip

The two young sea-lions took not the slightest interest

NEural Attention Tradeoff Model (NEAT)

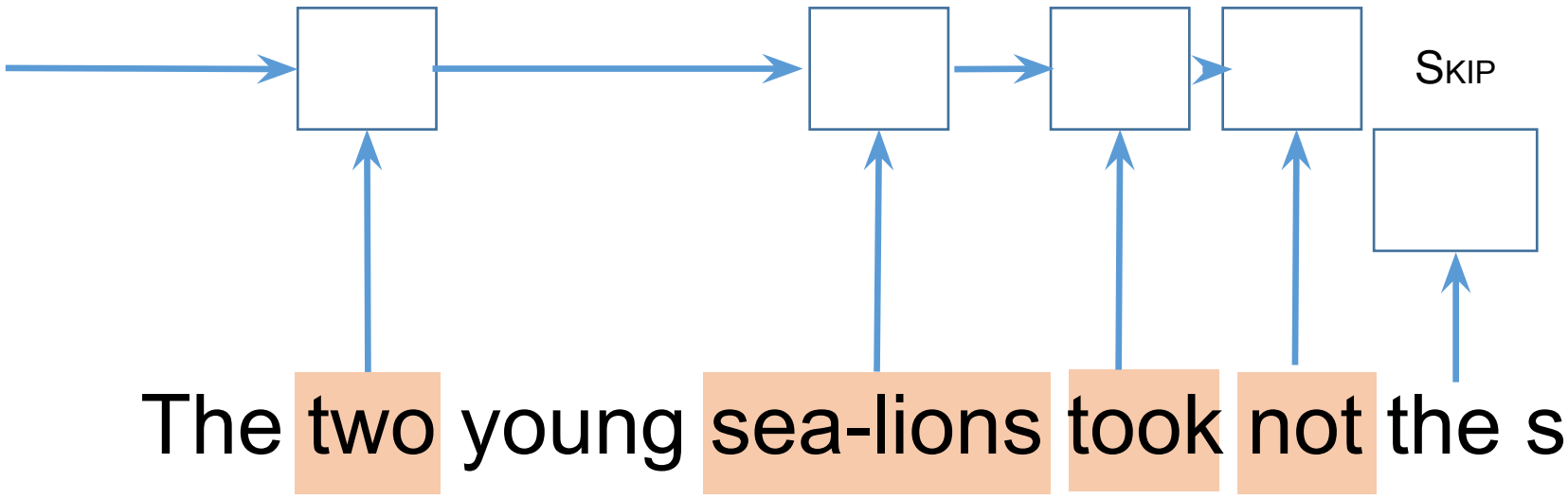
Reader Network
reads text



The two young sea-lions took not the slightest interest

NEural Attention Tradeoff Model (NEAT)

Reader Network
reads text

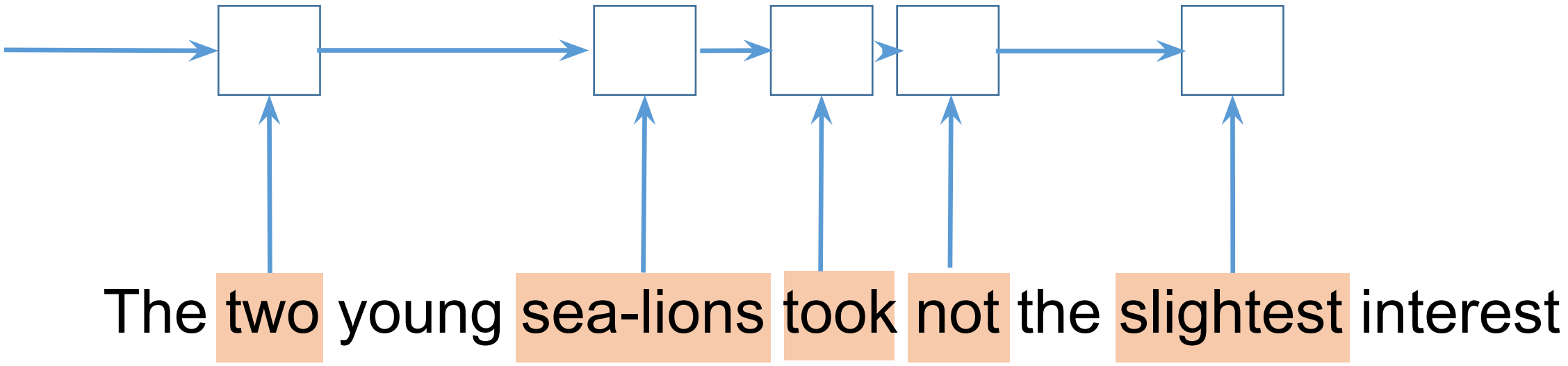


Control Network
decides which word to read or skip

The two young sea-lions took not the slightest interest

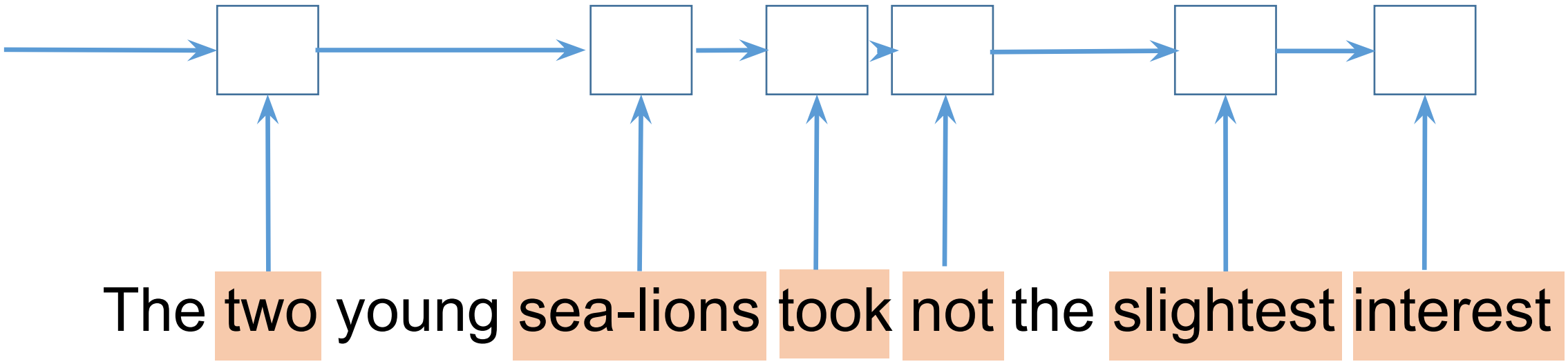
NEural Attention Tradeoff Model (NEAT)

Reader Network
reads text

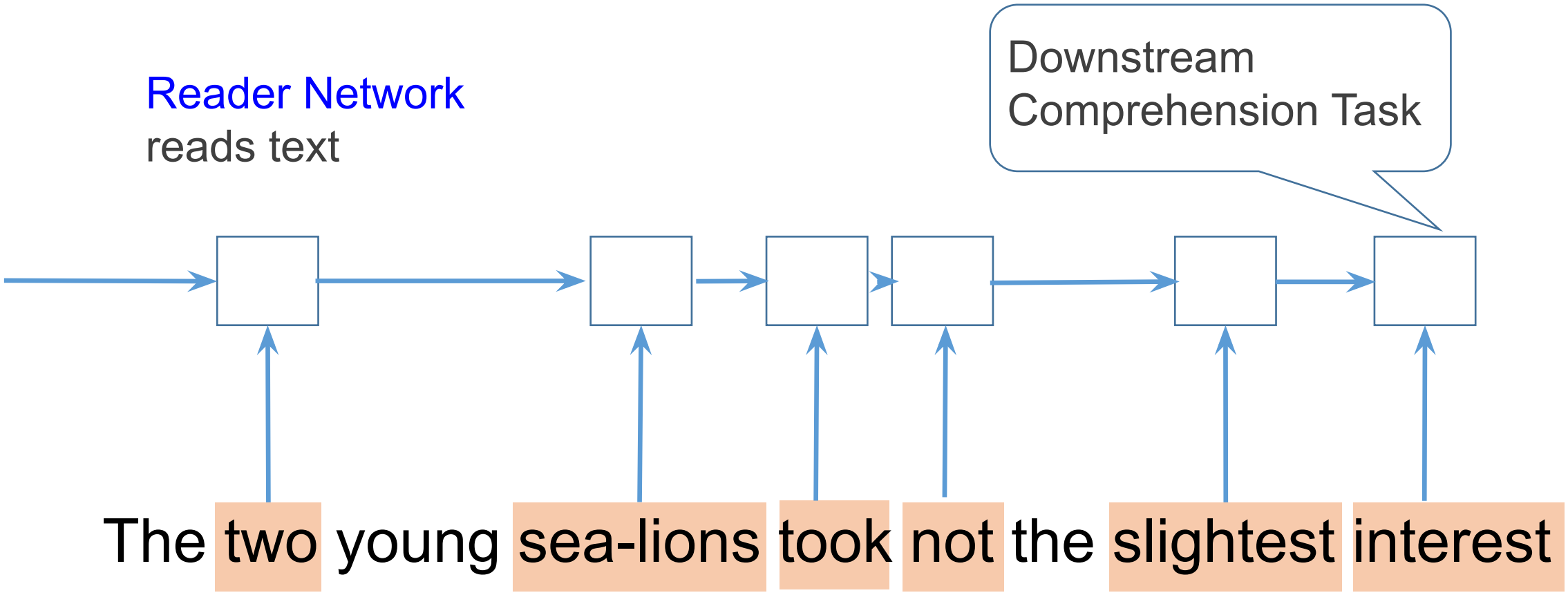


NEural Attention Tradeoff Model (NEAT)

Reader Network
reads text



NEural Attention Tradeoff Model (NEAT)



NEural Attention Tradeoff Model (NEAT)

Hypothesis: Human reading is optimized for

efficiently extracting information | while | maximizing speed

Training Objective

Solve downstream task with minimal attention:

$$\arg\min \mathbb{E}\,[\ \text{TaskLoss} + \alpha\ \#\{\text{Words Read}\}]$$

NEural Attention Tradeoff Model (NEAT)

Hypothesis: Human reading is optimized for

efficiently extracting information | while | maximizing speed

Training Objective

Solve downstream task with minimal attention:

$$\arg\min \mathbb{E}\,[\,\text{TaskLoss} + \alpha\ \#\{\text{Words Read}\}]$$

NEural Attention Tradeoff Model (NEAT)

Hypothesis: Human reading is optimized for

efficiently extracting information | while | maximizing speed

Training Objective

Solve downstream task with minimal attention:

$$\arg\min \mathbb{E}\,[\,\text{TaskLoss} + \alpha\,\#\{\text{Words Read}\}]$$

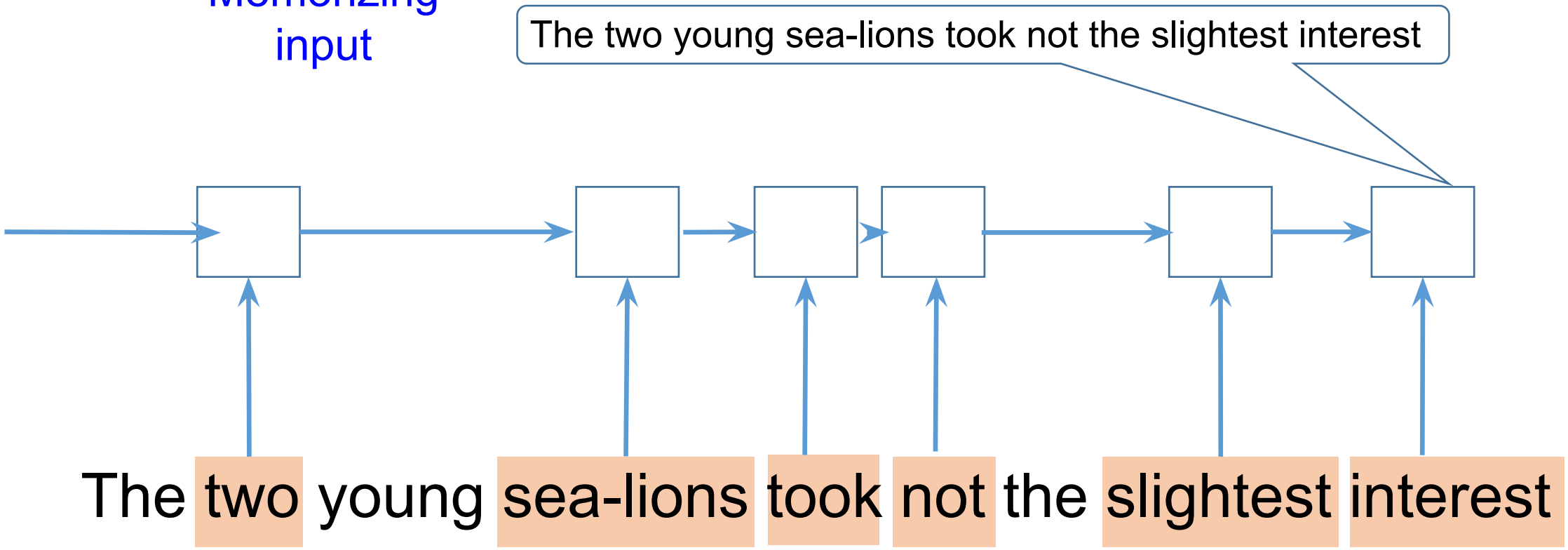
Trained using reinforcement learning (policy gradient).

NEural Attention Tradeoff Model (NEAT)

Task:

Memorizing input

TaskLoss = - log P(text|model output)

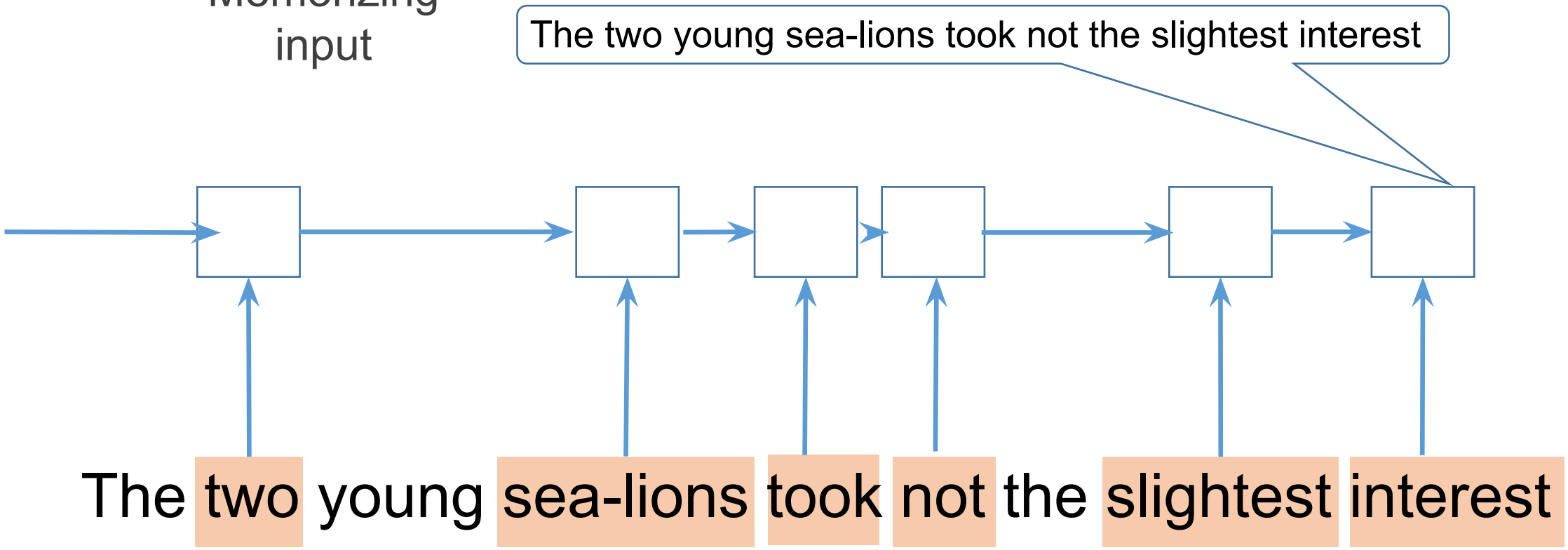


NEural Attention Tradeoff Model (NEAT)

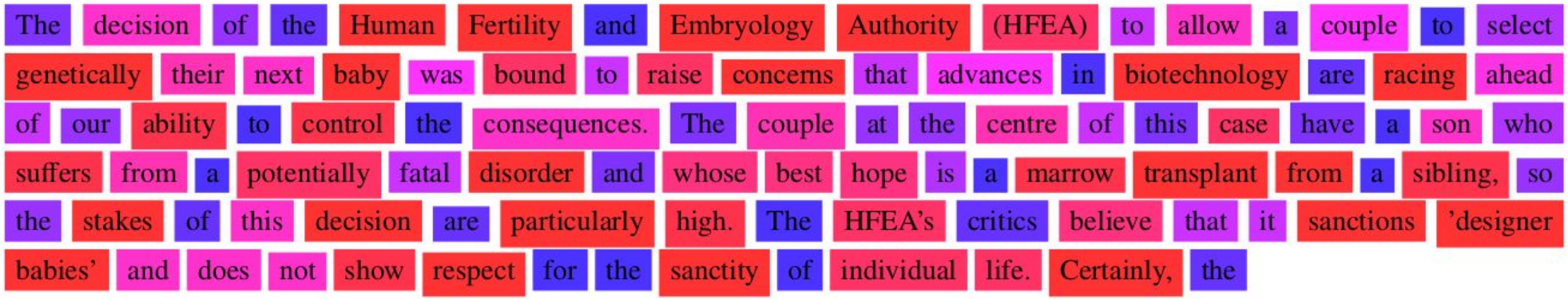
Task:

Memorizing input

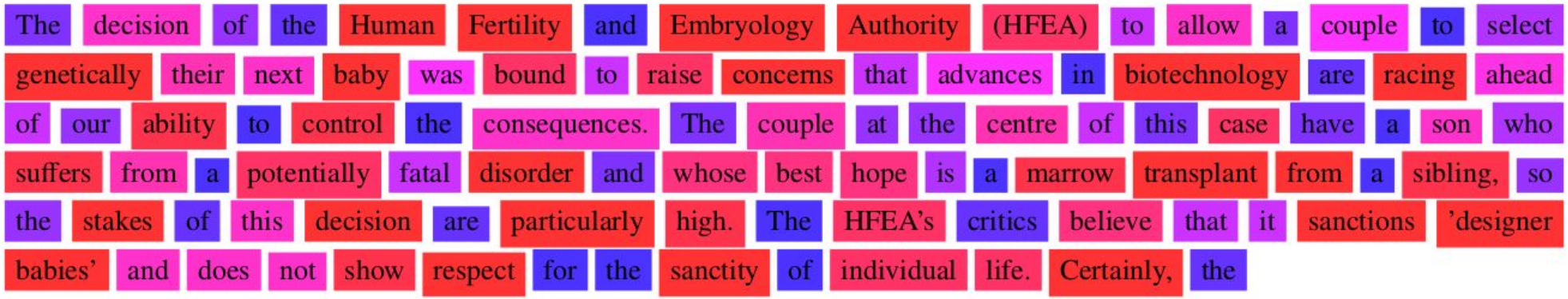
TaskLoss = - log P(text|model output)



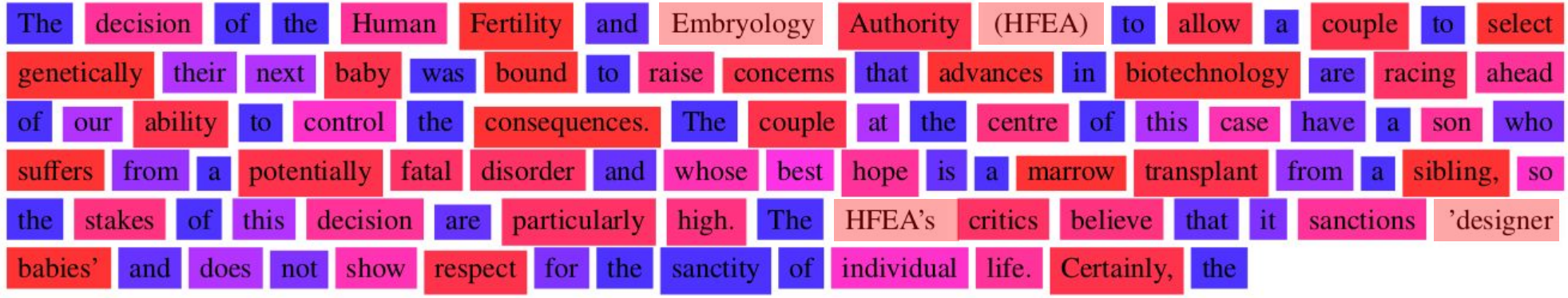
Humans


The decision of the Human Fertility and Embryology Authority (HFEA) to allow a couple to select
genetically their next baby was bound to raise concerns that advances in biotechnology are racing ahead
of our ability to control the consequences. The couple at the centre of this case have a son who
suffers from a potentially fatal disorder and whose best hope is a marrow transplant from a sibling, so
the stakes of this decision are particularly high. The HFEA's critics believe that it sanctions 'designer
babies' and does not show respect for the sanctity of individual life. Certainly, the

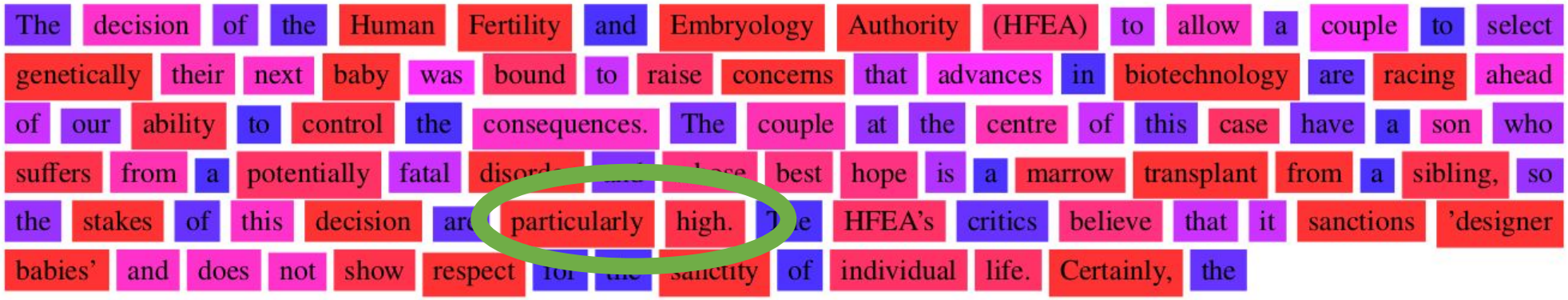
Humans



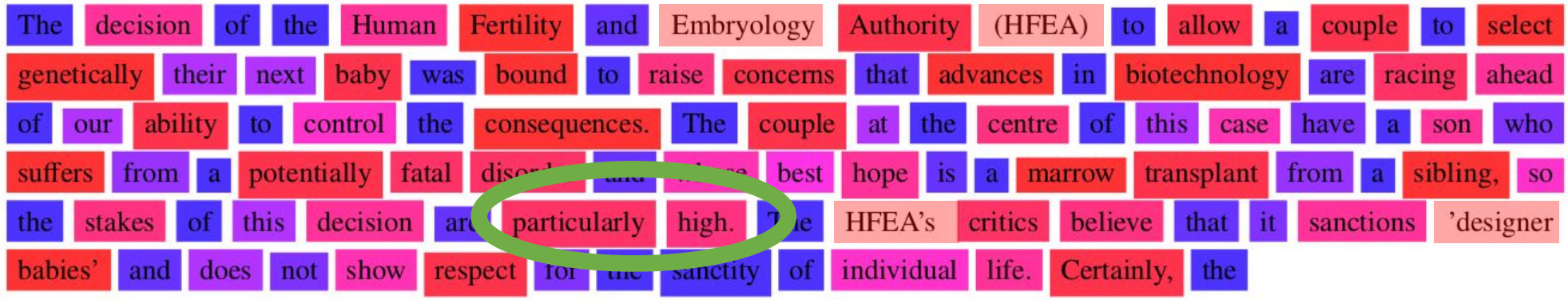
Model



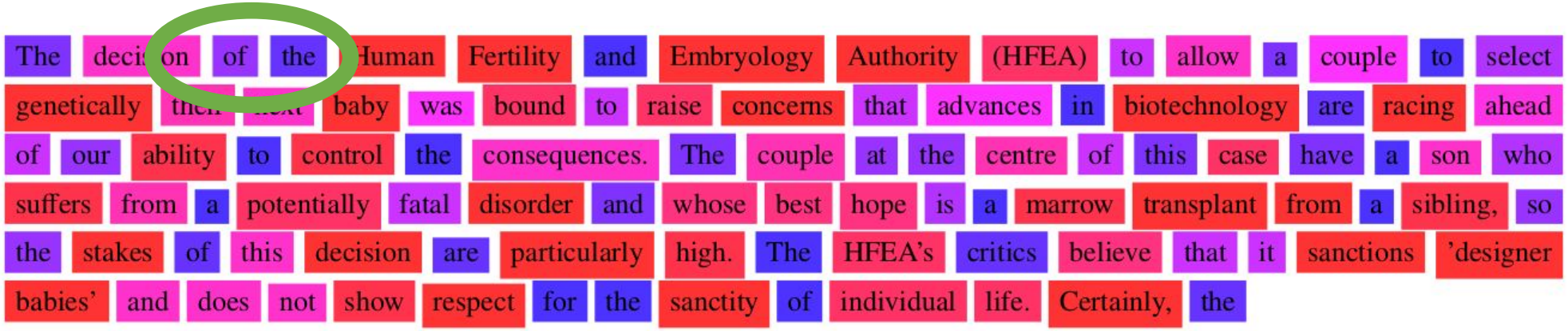
Humans



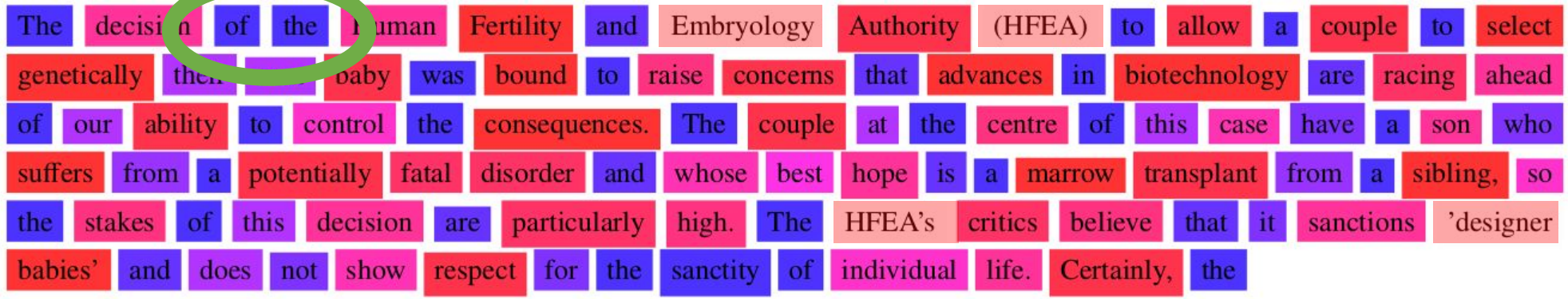
Model



Humans

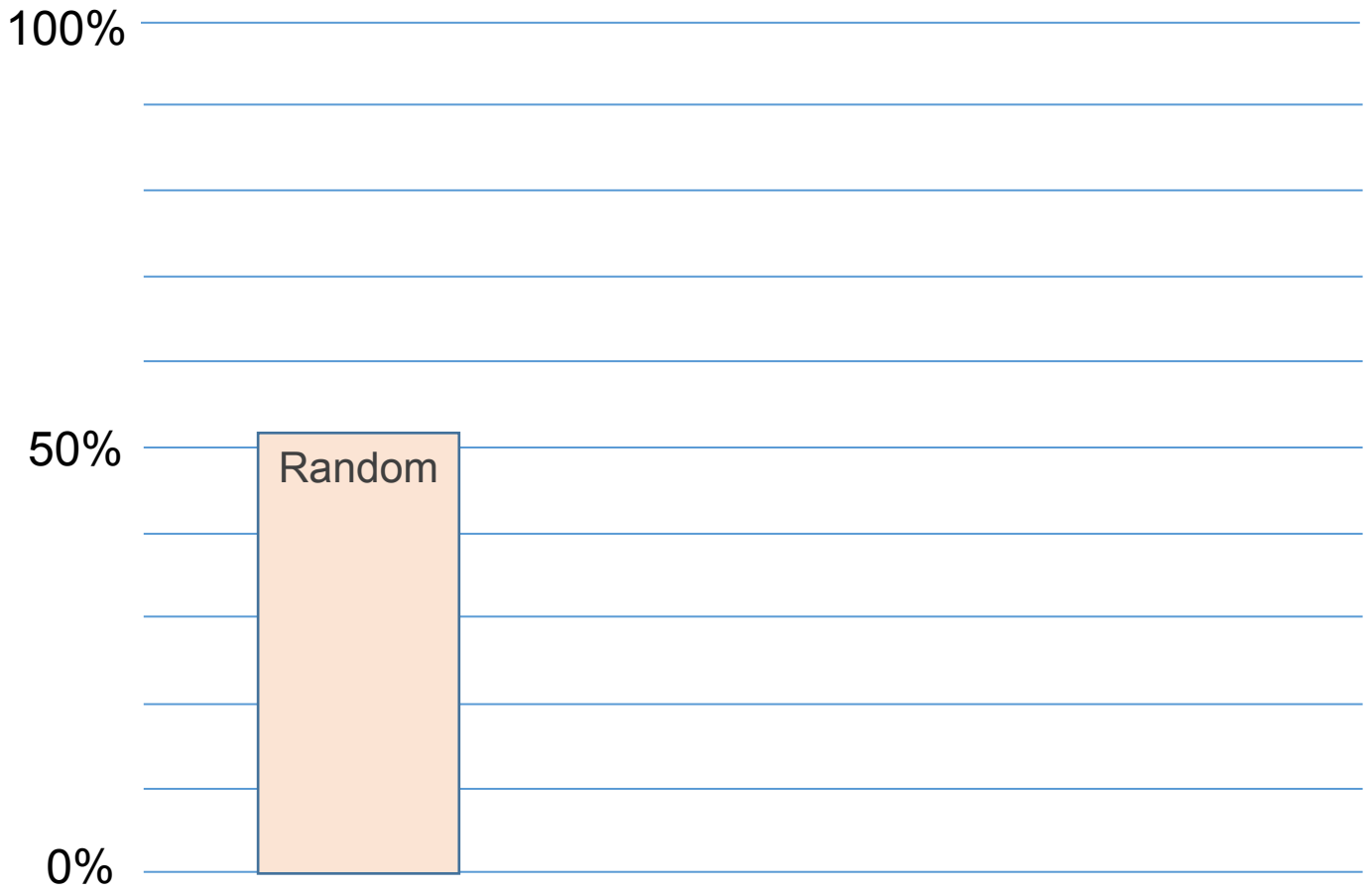


Model

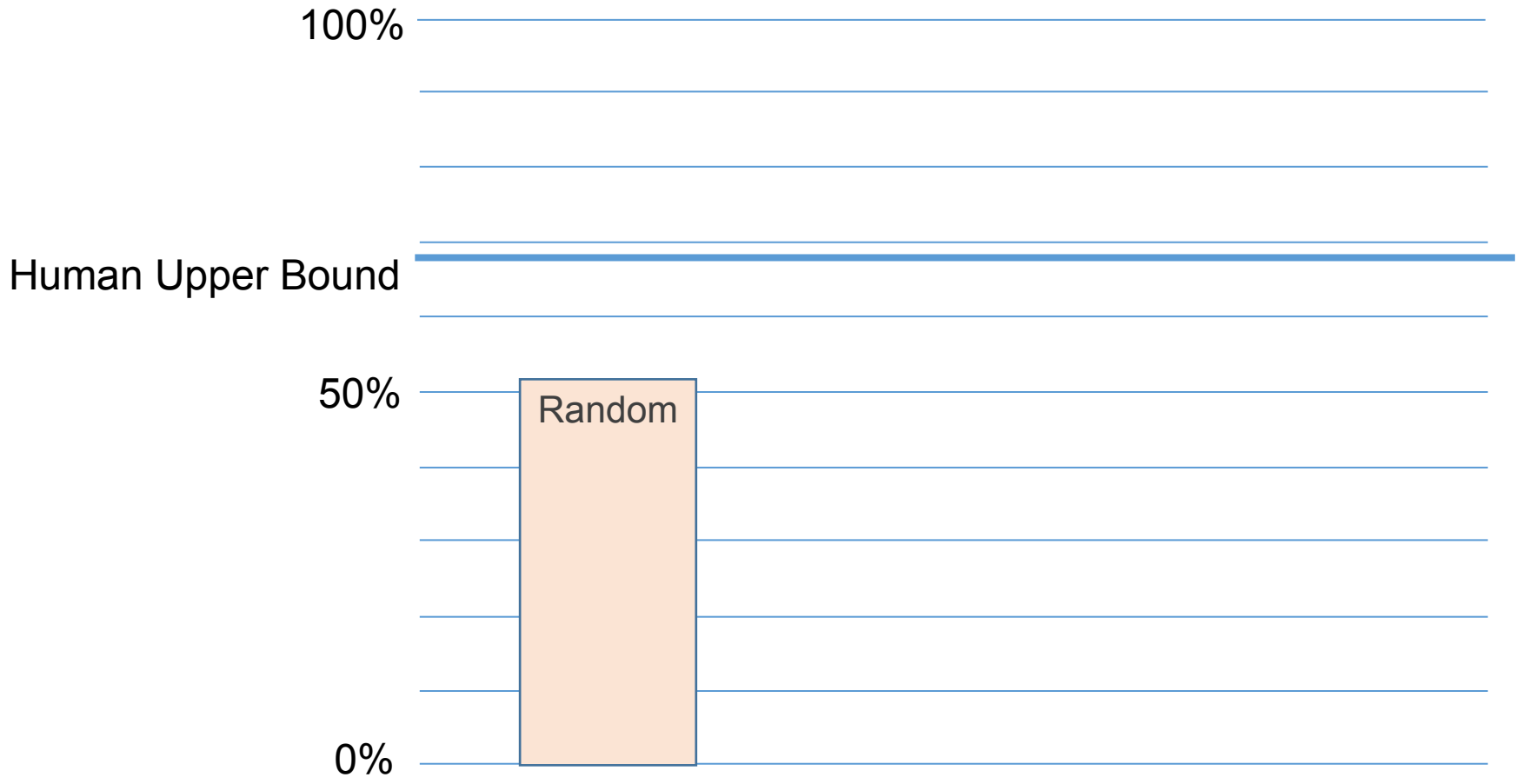


Accuracy Predicting Human Fixations

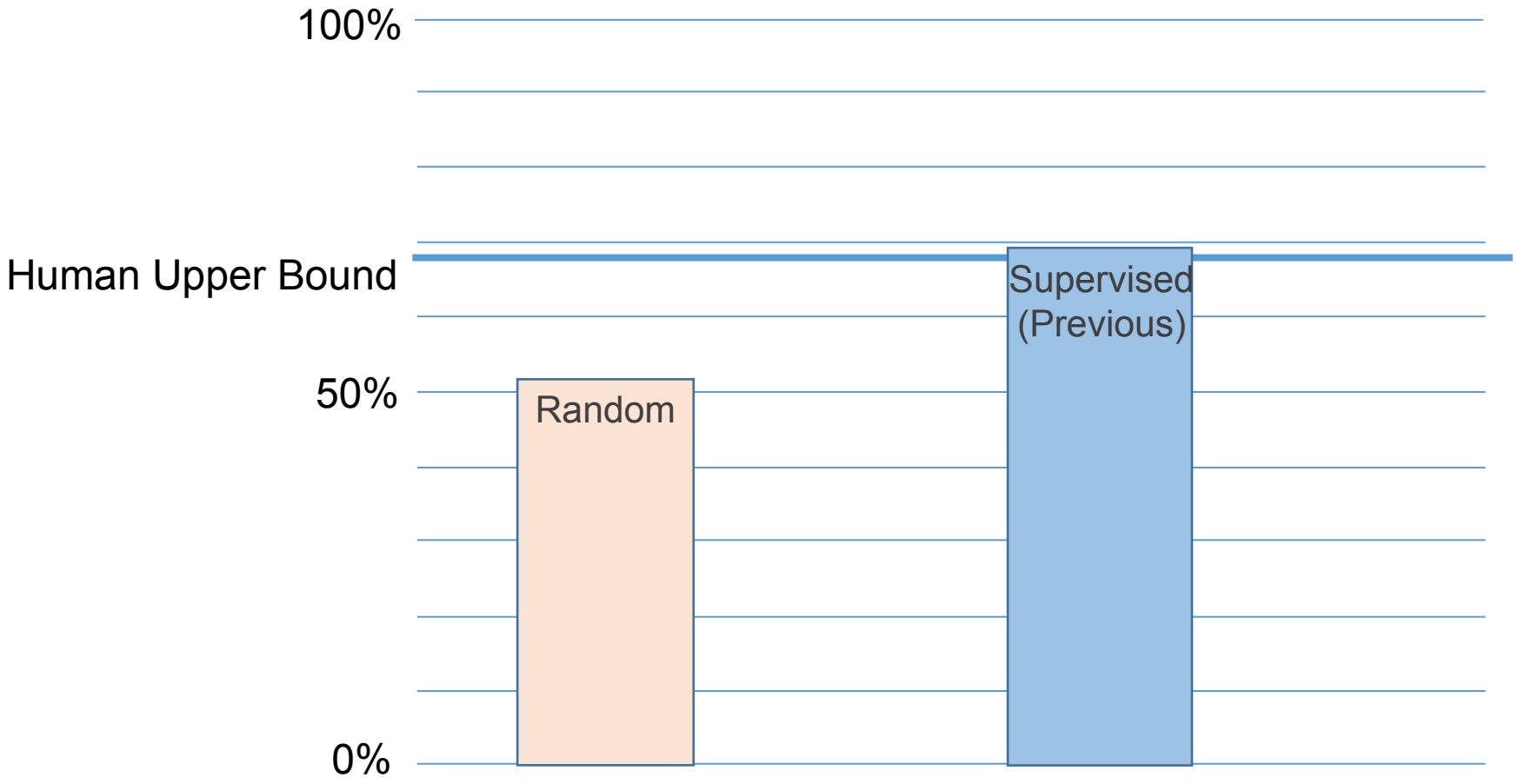
Accuracy Predicting Human Fixations



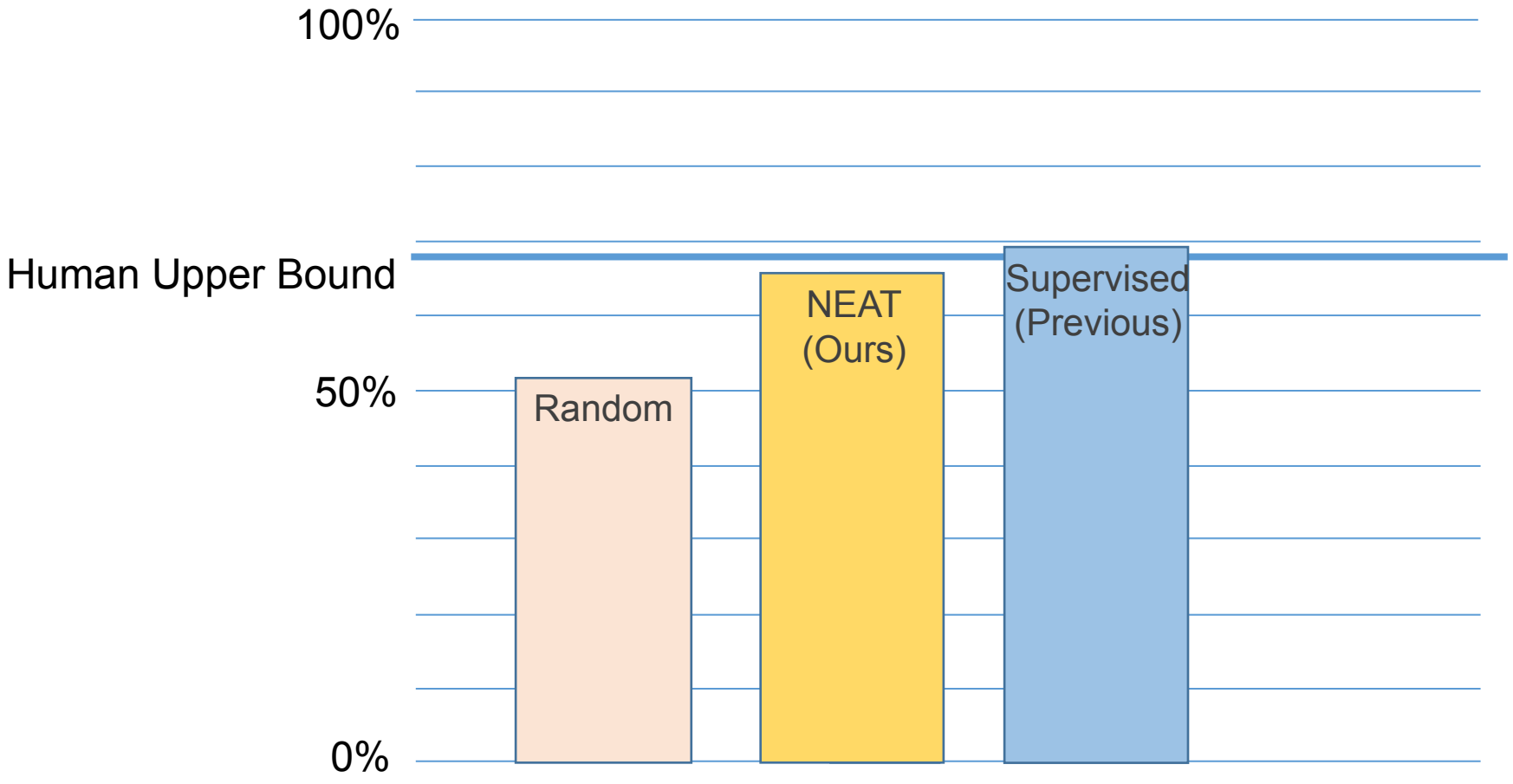
Accuracy Predicting Human Fixations



Accuracy Predicting Human Fixations



Accuracy Predicting Human Fixations



Human Eyetracking Experiment

Human Eyetracking Experiment

NoPreview Group (10 participants)

1

Sabra is recalling 30,000 cases of hummus due to possible contamination with Listeria, the U.S. said Wednesday. The nationwide recall is voluntary. So far, no illnesses caused by the hummus have been reported. The potential for contamination was discovered when a routine, random sample collected at a Michigan store on March 30 tested positive for Listeria monocytogenes. The FDA issued a list of the products in the recall. Anyone who has purchased any of the items is urged to dispose of or return it to the store for a full refund. Listeria monocytogenes can cause serious and sometimes fatal infections in young children, frail or elderly people, and others with weakened immune systems, the FDA says. Although some people may suffer only short-term symptoms such as high fever, severe headache, nausea, abdominal pain and diarrhea, Listeria can also cause miscarriages and stillbirths among pregnant women.

Human Eyetracking Experiment

NoPreview Group (10 participants)

1

Sabra is recalling 30,000 cases of hummus due to possible contamination with Listeria, the U.S. said Wednesday. The nationwide recall is voluntary. So far, no illnesses caused by the hummus have been reported. The potential for contamination was discovered when a routine, random sample collected at a Michigan store on March 30 tested positive for Listeria monocytogenes. The FDA issued a list of the products in the recall. Anyone who has purchased any of the items is urged to dispose of or return it to the store for a full refund. Listeria monocytogenes can cause serious and sometimes fatal infections in young children, frail or elderly people, and others with weakened immune systems, the FDA says. Although some people may suffer only short-term symptoms such as high fever, severe headache, nausea, abdominal pain and diarrhea, Listeria can also cause miscarriages and stillbirths among pregnant women.

2

Question: A random sample from a _____ store tested positive for Listeria monocytogenes.

Answers: (1) Michigan (2) Washington (3) Ohio (4) Georgia

Human Eyetracking Experiment

NoPreview Group (10 participants)

1

Sabra is recalling 30,000 cases of hummus due to possible contamination with Listeria, the U.S. said Wednesday. The nationwide recall is voluntary. So far, no illnesses caused by the hummus have been reported. The potential for contamination was discovered when a routine, random sample collected at a Michigan store on March 30 tested positive for Listeria monocytogenes. The FDA issued a list of the products in the recall. Anyone who has purchased any of the items is urged to dispose of or return it to the store for a full refund. Listeria monocytogenes can cause serious and sometimes fatal infections in young children, frail or elderly people, and others with weakened immune systems, the FDA says. Although some people may suffer only short-term symptoms such as high fever, severe headache, nausea, abdominal pain and diarrhea, Listeria can also cause miscarriages and stillbirths among pregnant women.

2

Question: A random sample from a _____ store tested positive for Listeria monocytogenes.

Answers: (1) Michigan (2) Washington (3) Ohio (4) Georgia

WithPreview Group (10 participants)

1

Question: A random sample from a _____ store tested positive for Listeria monocytogenes.

Human Eyetracking Experiment

NoPreview Group (10 participants)

1

Sabra is recalling 30,000 cases of hummus due to possible contamination with Listeria, the U.S. said Wednesday. The nationwide recall is voluntary. So far, no illnesses caused by the hummus have been reported. The potential for contamination was discovered when a routine, random sample collected at a Michigan store on March 30 tested positive for Listeria monocytogenes. The FDA issued a list of the products in the recall. Anyone who has purchased any of the items is urged to dispose of or return it to the store for a full refund. Listeria monocytogenes can cause serious and sometimes fatal infections in young children, frail or elderly people, and others with weakened immune systems, the FDA says. Although some people may suffer only short-term symptoms such as high fever, severe headache, nausea, abdominal pain and diarrhea, Listeria can also cause miscarriages and stillbirths among pregnant women.

2

Question: A random sample from a _____ store tested positive for Listeria monocytogenes.

Answers: (1) Michigan (2) Washington (3) Ohio (4) Georgia

WithPreview Group (10 participants)

1

Question: A random sample from a _____ store tested positive for Listeria monocytogenes.

2

Sabra is recalling 30,000 cases of hummus due to possible contamination with Listeria, the U.S. said Wednesday. The nationwide recall is voluntary. So far, no illnesses caused by the hummus have been reported. The potential for contamination was discovered when a routine, random sample collected at a Michigan store on March 30 tested positive for Listeria monocytogenes. The FDA issued a list of the products in the recall. Anyone who has purchased any of the items is urged to dispose of or return it to the store for a full refund. Listeria monocytogenes can cause serious and sometimes fatal infections in young children, frail or elderly people, and others with weakened immune systems, the FDA says. Although some people may suffer only short-term symptoms such as high fever, severe headache, nausea, abdominal pain and diarrhea, Listeria can also cause miscarriages and stillbirths among pregnant women.

Human Eyetracking Experiment

NoPreview Group (10 participants)

1

Sabra is recalling 30,000 cases of hummus due to possible contamination with Listeria, the U.S. said Wednesday. The nationwide recall is voluntary. So far, no illnesses caused by the hummus have been reported. The potential for contamination was discovered when a routine, random sample collected at a Michigan store on March 30 tested positive for Listeria monocytogenes. The FDA issued a list of the products in the recall. Anyone who has purchased any of the items is urged to dispose of or return it to the store for a full refund. Listeria monocytogenes can cause serious and sometimes fatal infections in young children, frail or elderly people, and others with weakened immune systems, the FDA says. Although some people may suffer only short-term symptoms such as high fever, severe headache, nausea, abdominal pain and diarrhea, Listeria can also cause miscarriages and stillbirths among pregnant women.

2

Question: A random sample from a _____ store tested positive for Listeria monocytogenes.

Answers: (1) Michigan (2) Washington (3) Ohio (4) Georgia

WithPreview Group (10 participants)

1

Question: A random sample from a _____ store tested positive for Listeria monocytogenes.

2

Sabra is recalling 30,000 cases of hummus due to possible contamination with Listeria, the U.S. said Wednesday. The nationwide recall is voluntary. So far, no illnesses caused by the hummus have been reported. The potential for contamination was discovered when a routine, random sample collected at a Michigan store on March 30 tested positive for Listeria monocytogenes. The FDA issued a list of the products in the recall. Anyone who has purchased any of the items is urged to dispose of or return it to the store for a full refund. Listeria monocytogenes can cause serious and sometimes fatal infections in young children, frail or elderly people, and others with weakened immune systems, the FDA says. Although some people may suffer only short-term symptoms such as high fever, severe headache, nausea, abdominal pain and diarrhea, Listeria can also cause miscarriages and stillbirths among pregnant women.

3

Question: A random sample from a _____ store tested positive for Listeria monocytogenes.

Answers: (1) Michigan (2) Washington (3) Ohio (4) Georgia

Human Eyetracking Experiment

NoPreview Group (10 participants)

1

Sabra is recalling 30,000 cases of hummus due to possible contamination with Listeria, the U.S. said Wednesday. The nationwide recall is voluntary. So far, no illnesses caused by the hummus have been reported. The potential for contamination was discovered when a routine, random sample collected at a Michigan store on March 30 tested positive for Listeria monocytogenes. The FDA issued a list of the products in the recall. Anyone who has purchased any of the items is urged to dispose of or return it to the store for a full refund. Listeria monocytogenes can cause serious and sometimes fatal infections in young children, frail or elderly people, and others with weakened immune systems, the FDA says. Although some people may suffer only short-term symptoms such as high fever, severe headache, nausea, abdominal pain and diarrhea, Listeria can also cause miscarriages and stillbirths among pregnant women.

2

Question: A random sample from a _____ store tested positive for Listeria monocytogenes.

Answers: (1) Michigan (2) Washington (3) Ohio (4) Georgia

WithPreview Group (10 participants)

1

Question: A random sample from a _____ store tested positive for Listeria monocytogenes.

2

Sabra is recalling 30,000 cases of hummus due to possible contamination with Listeria, the U.S. said Wednesday. The nationwide recall is voluntary. So far, no illnesses caused by the hummus have been reported. The potential for contamination was discovered when a routine, random sample collected at a Michigan store on March 30 tested positive for Listeria monocytogenes. The FDA issued a list of the products in the recall. Anyone who has purchased any of the items is urged to dispose of or return it to the store for a full refund. Listeria monocytogenes can cause serious and sometimes fatal infections in young children, frail or elderly people, and others with weakened immune systems, the FDA says. Although some people may suffer only short-term symptoms such as high fever, severe headache, nausea, abdominal pain and diarrhea, Listeria can also cause miscarriages and stillbirths among pregnant women.

3

Question: A random sample from a _____ store tested positive for Listeria monocytogenes.

Answers: (1) Michigan (2) Washington (3) Ohio (4) Georgia

Human Eyetracking Experiment

NoPreview Group (10 participants)

1

Sabra is recalling 30,000 cases of hummus due to possible contamination with Listeria, the U.S. said Wednesday. The nationwide recall is voluntary. So far, no illnesses caused by the hummus have been reported. The potential for contamination was discovered when a routine, random sample collected at a Michigan store on March 30 tested positive for Listeria monocytogenes. The FDA issued a list of the products in the recall. Anyone who has purchased any of the items is urged to dispose of or return it to the store for a full refund. Listeria monocytogenes can cause serious and sometimes fatal infections in young children, frail or elderly people, and others with weakened immune systems, the FDA says. Although some people may suffer only short-term symptoms such as high fever, severe headache, nausea, abdominal pain and diarrhea, Listeria can also cause miscarriages and stillbirths among pregnant women.

2

Question: A random sample from a _____ store tested positive for Listeria monocytogenes.

Answers: (1) Michigan (2) Washington (3) Ohio (4) Georgia

WithPreview Group (10 participants)

1

Question: A random sample from a _____ store tested positive for Listeria monocytogenes.

2

Sabra is recalling 30,000 cases of hummus due to possible contamination with Listeria, the U.S. said Wednesday. The nationwide recall is voluntary. So far, no illnesses caused by the hummus have been reported. The potential for contamination was discovered when a routine, random sample collected at a Michigan store on March 30 tested positive for Listeria monocytogenes. The FDA issued a list of the products in the recall. Anyone who has purchased any of the items is urged to dispose of or return it to the store for a full refund. Listeria monocytogenes can cause serious and sometimes fatal infections in young children, frail or elderly people, and others with weakened immune systems, the FDA says. Although some people may suffer only short-term symptoms such as high fever, severe headache, nausea, abdominal pain and diarrhea, Listeria can also cause miscarriages and stillbirths among pregnant women.

3

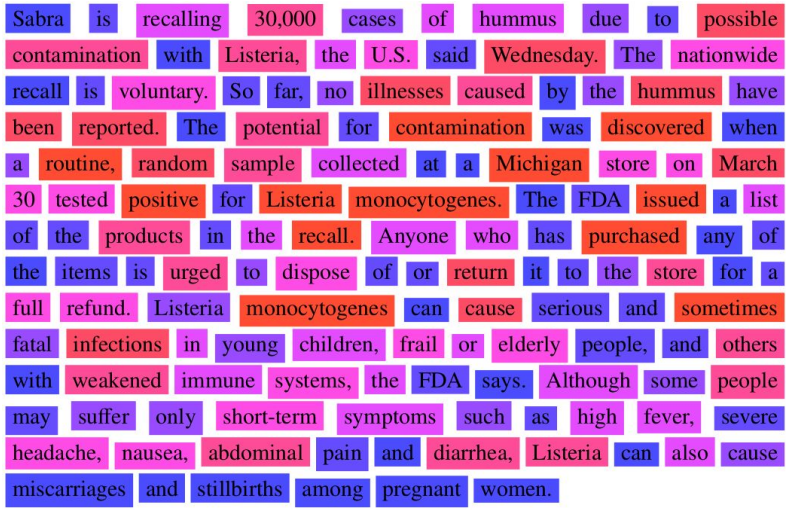
Question: A random sample from a _____ store tested positive for Listeria monocytogenes.

Answers: (1) Michigan (2) Washington (3) Ohio (4) Georgia

Human Eyetracking Experiment

NoPreview Group (10 participants)

WithPreview Group (10 participants)

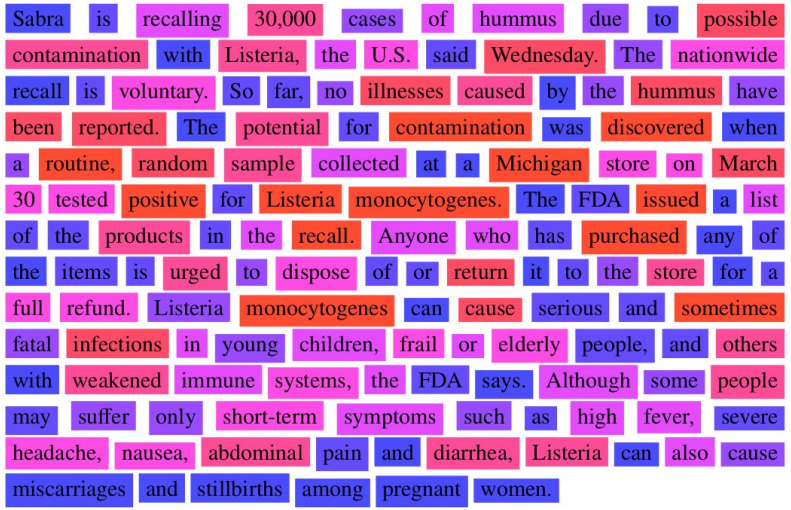


Human Eyetracking Experiment

NoPreview Group (10 participants)

More fixations overall

WithPreview Group (10 participants)



Human Eyetracking Experiment

NoPreview Group (10 participants)

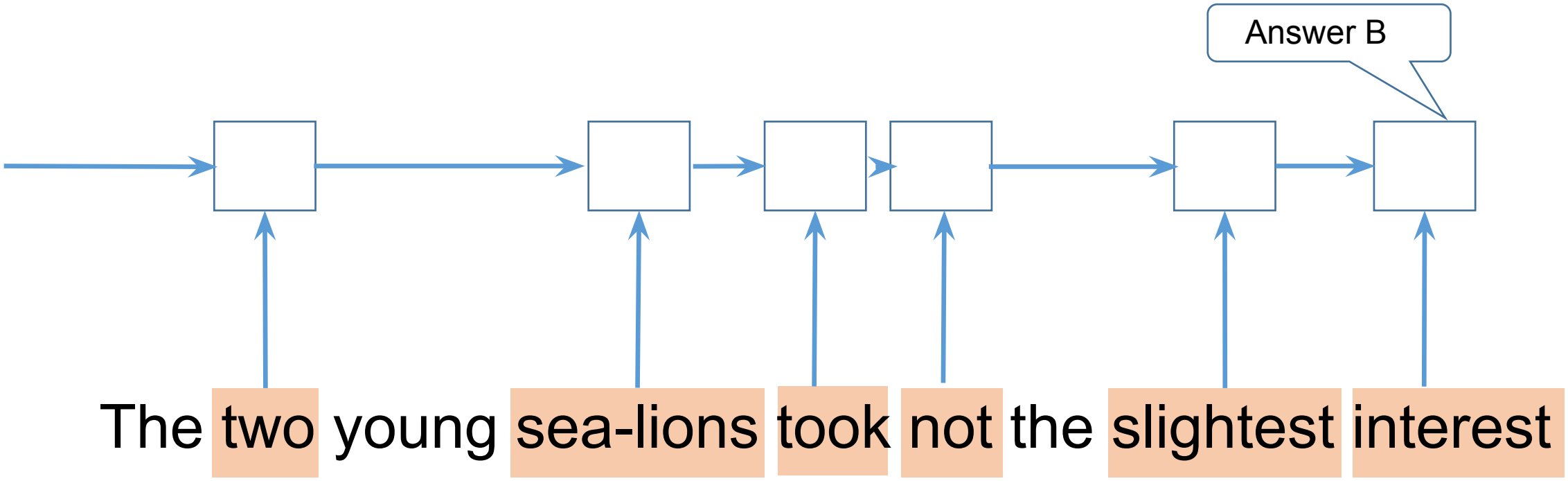
More fixations overall

WithPreview Group (10 participants)

More focused reading

NEural Attention Tradeoff Model (NEAT)

Task Loss: -log P(answer)



Model Behavior

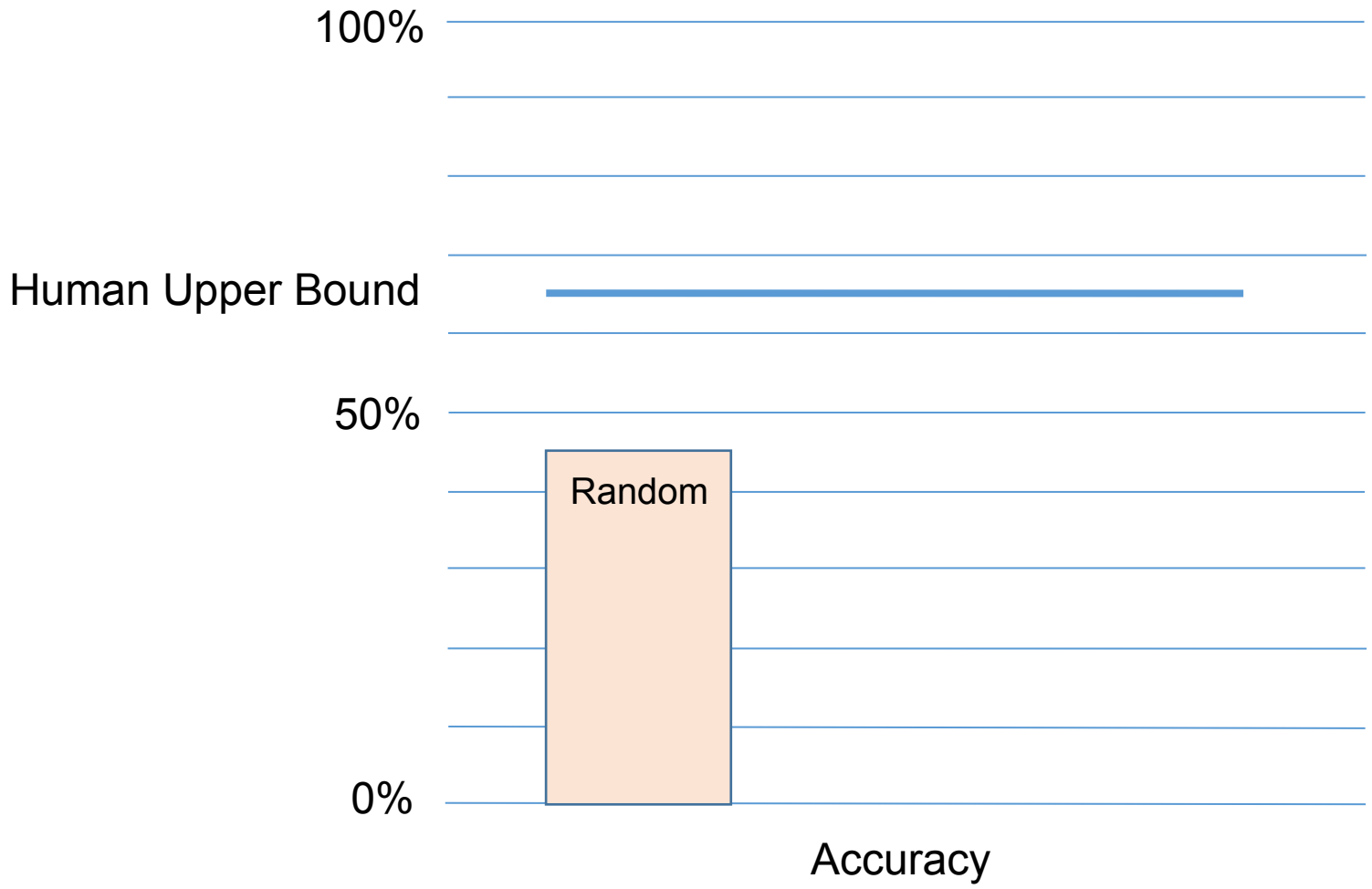
NoPreview Group (Model)

More fixations overall

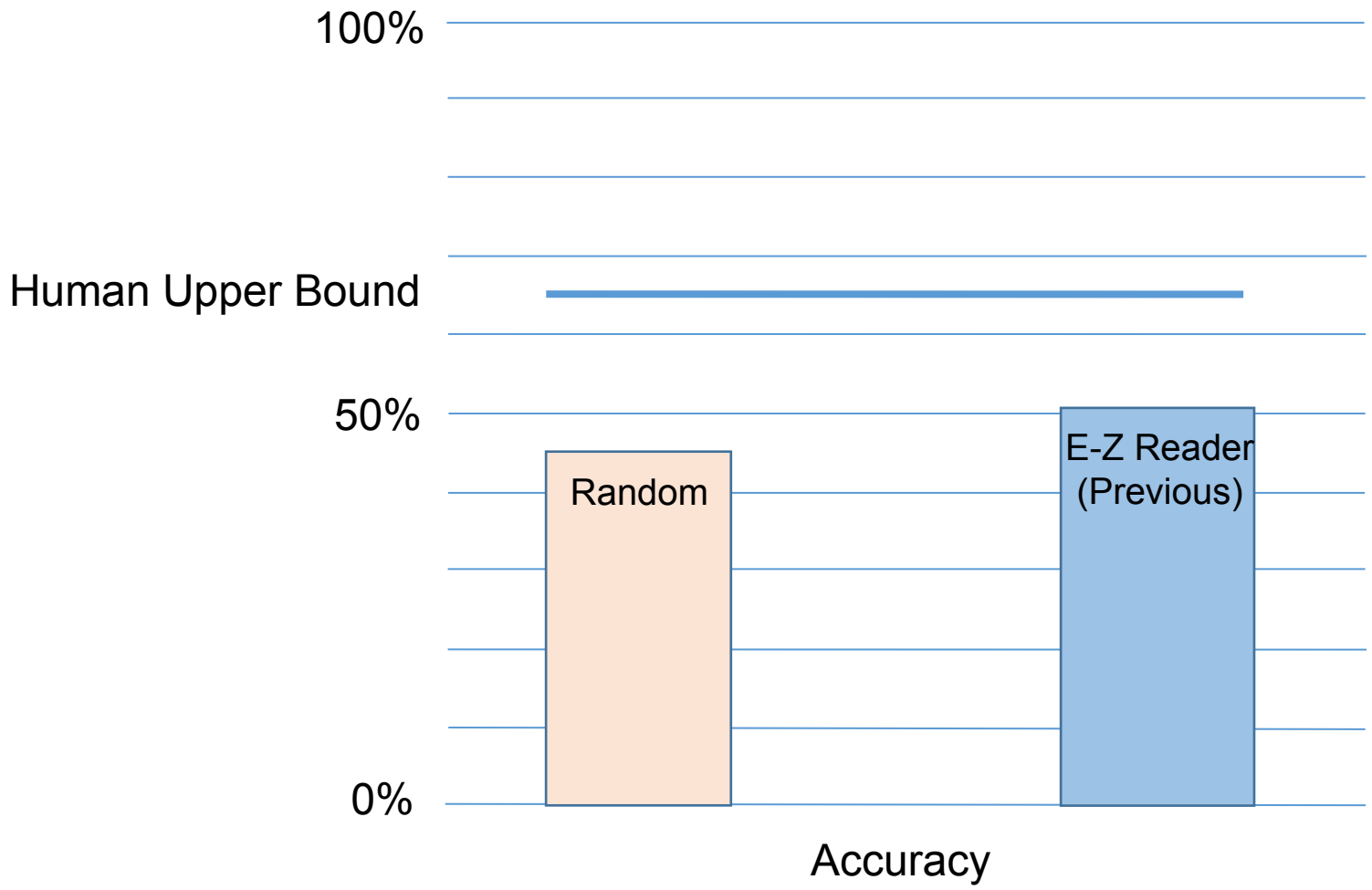
WithPreview Group (Model)

More focused reading

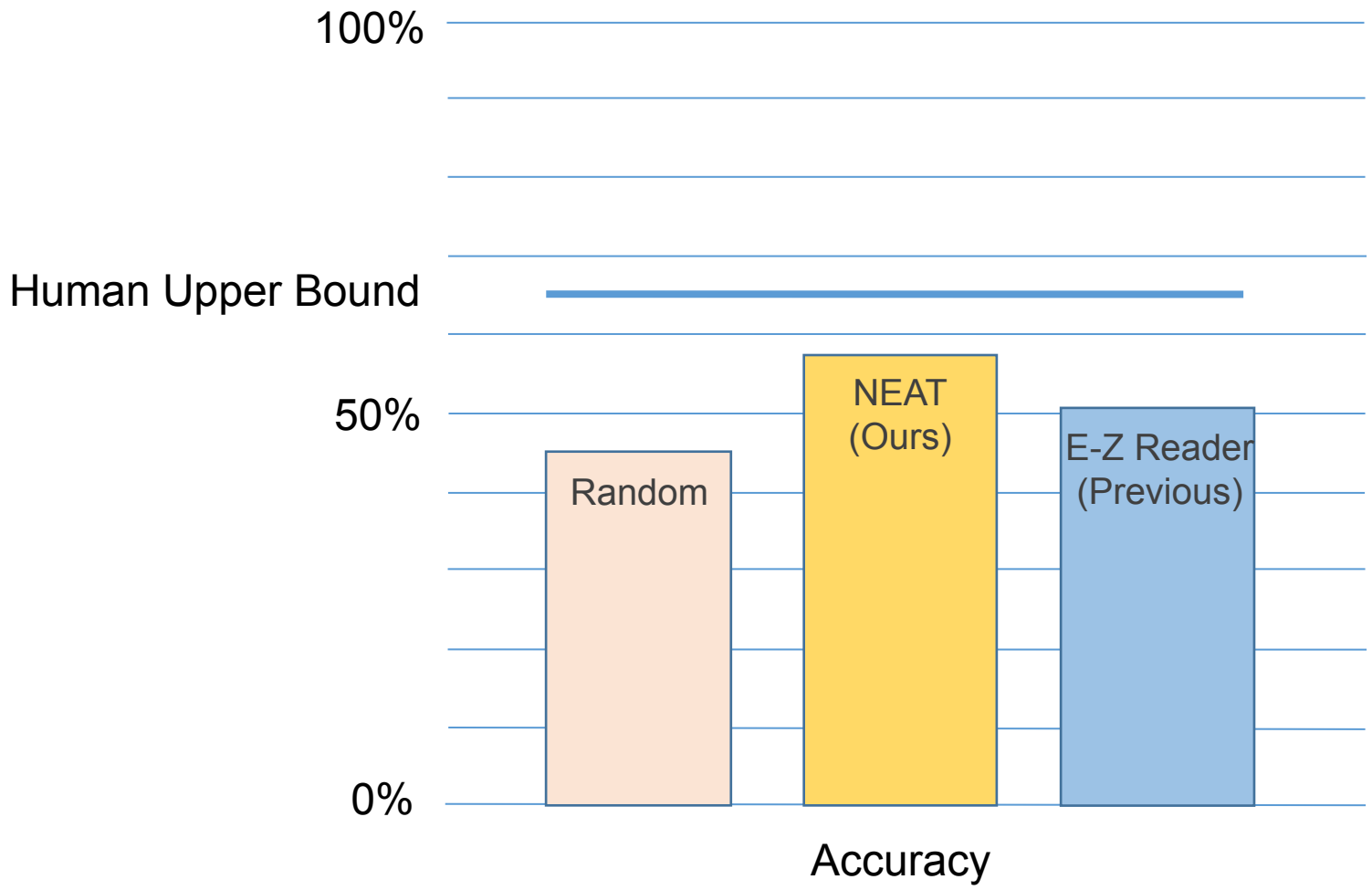
Accuracy Predicting Human Fixations



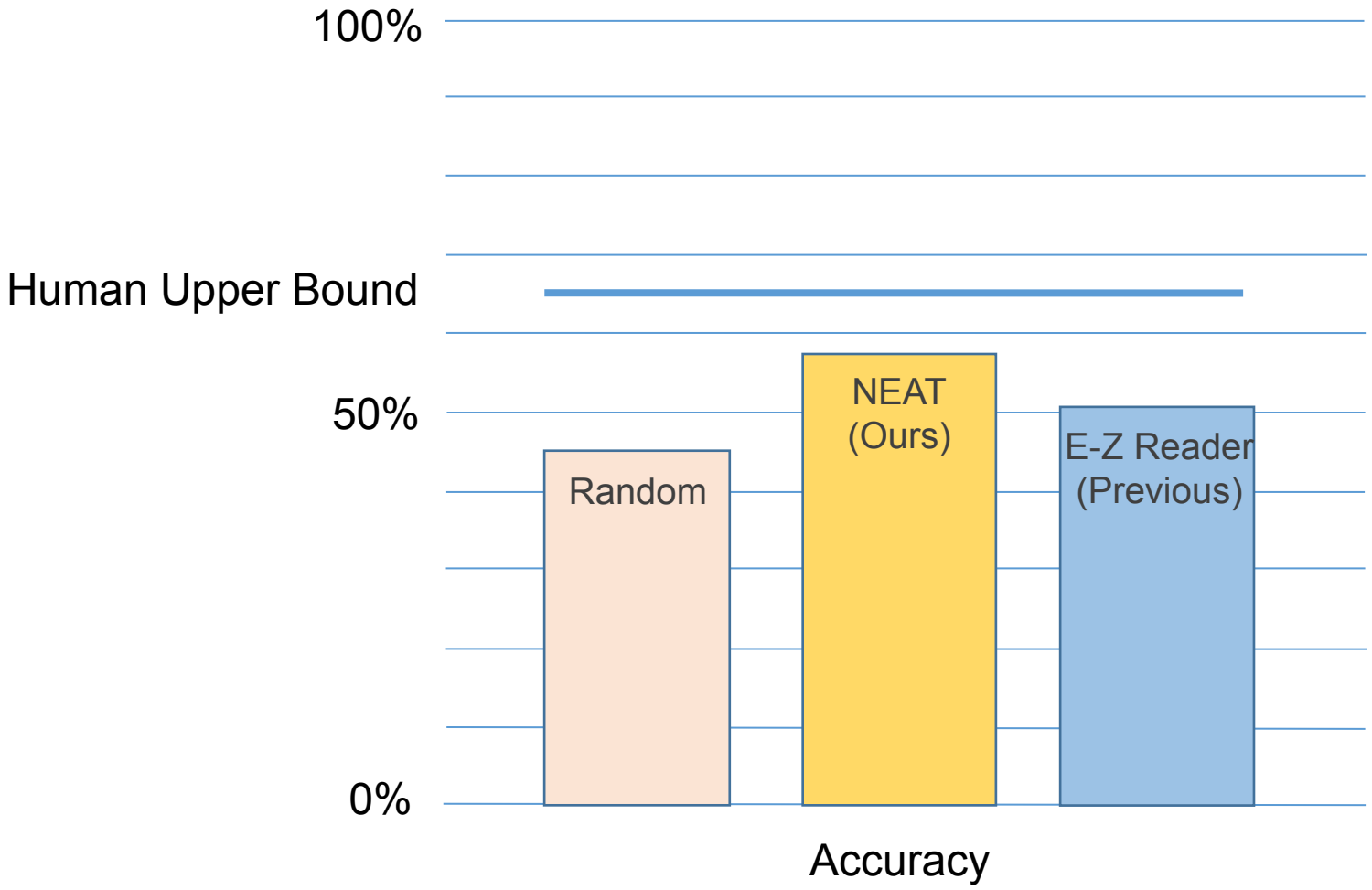
Accuracy Predicting Human Fixations



Accuracy Predicting Human Fixations



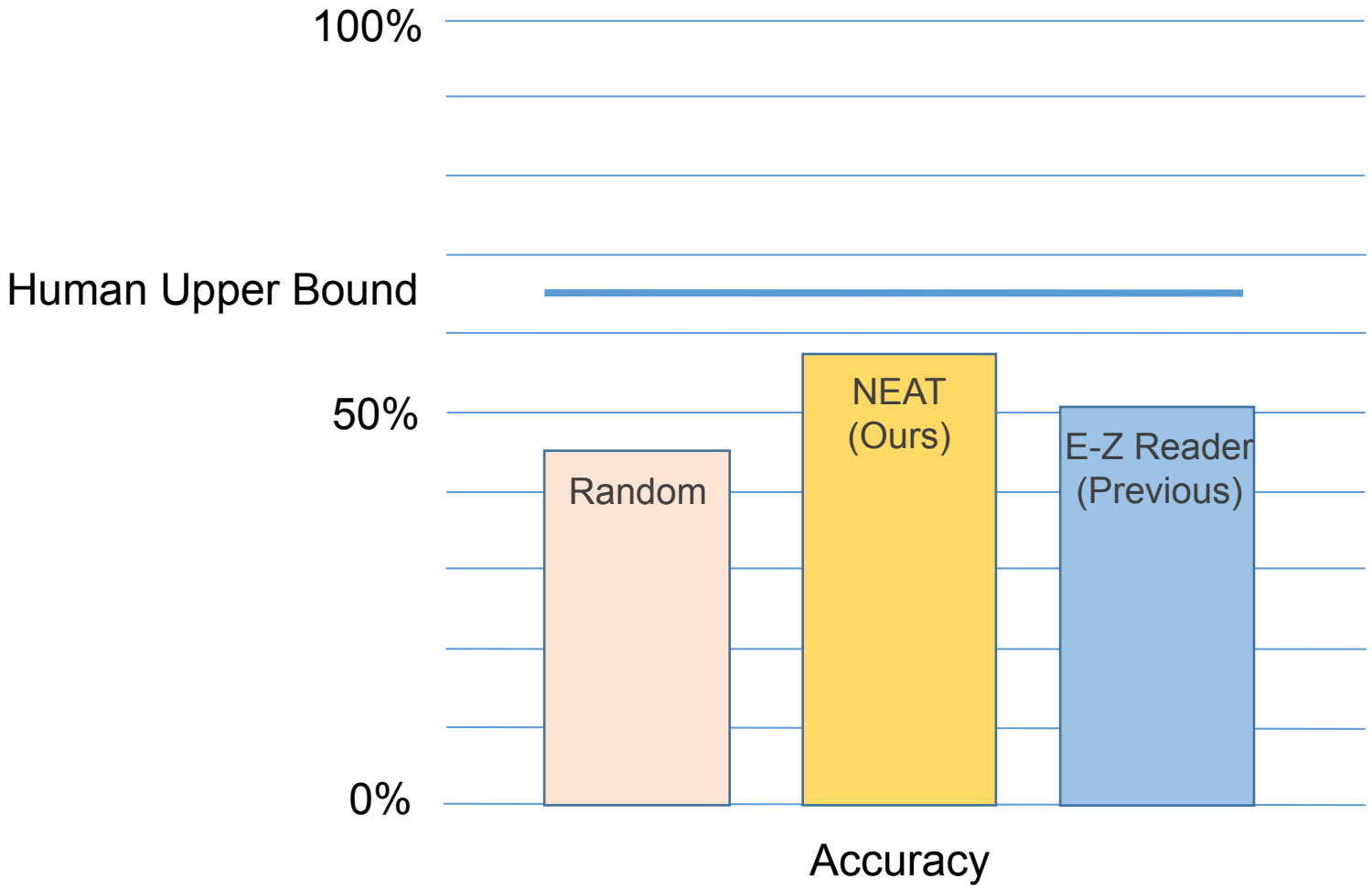
Accuracy Predicting Human Fixations



Take-Away:

Unsupervised machine learning model matches human reading better than previous models

Accuracy Predicting Human Fixations



Take-Away:

Unsupervised machine learning model matches human reading better than previous models

Points towards ways of making machine reading more human-like

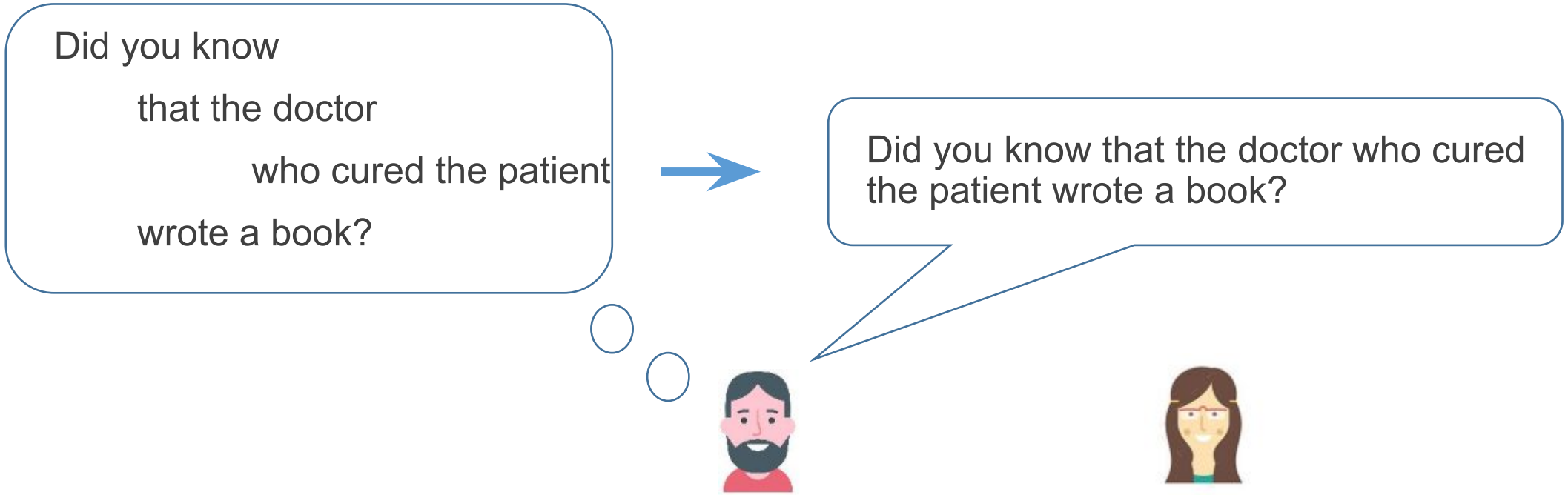
Motivation

Language expresses recursive thoughts via linear strings of words.

Did you know
that the doctor
who cured the patient
wrote a book?

Motivation

Language expresses recursive thoughts via linear strings of words.



Motivation

We have to recover hierarchical structure from linear sequences on a daily basis.

Did you know

that the doctor

who cured the patient

wrote a book?

Did you know that the doctor who cured the patient wrote a book?

Motivation

We have to recover hierarchical structure from linear sequences on a daily basis.

Did you know

that the doctor

who cured the patient

wrote a book?

Did you know that the doctor who cured the patient wrote a book?

But this can sometimes be quite difficult for our minds.

The

report

was

surprising

.
?

The

report

that

the

doctor

who

the

lawyer

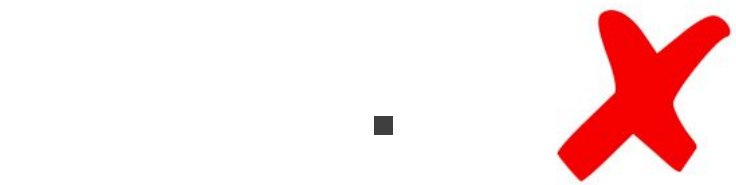
distrusted

annoyed

the

patient

.
?



The report that the doctor who the lawyer distrusted annoyed the patient.

The report

[that the doctor

[who the lawyer distrusted]

annoyed the patient].

The report

[that the doctor

[who the lawyer distrusted]

annoyed the patient]

?????.

The report

[that the doctor

[who the lawyer distrusted]

annoyed the patient]

was surprising.

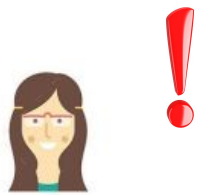
The report

[that the doctor

[who the lawyer distrusted]

annoyed the patient]

was surprising.



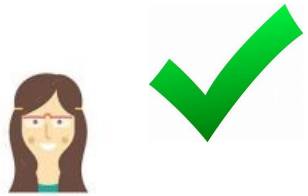
The **fact**

[that the doctor

[who the lawyer distrusted]

annoyed the patient]

was surprising.



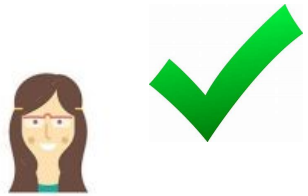
The **fact**

[that the doctor

[who the lawyer distrusted]

cured the patient]

was surprising.



Comprehending Hierarchical Structure

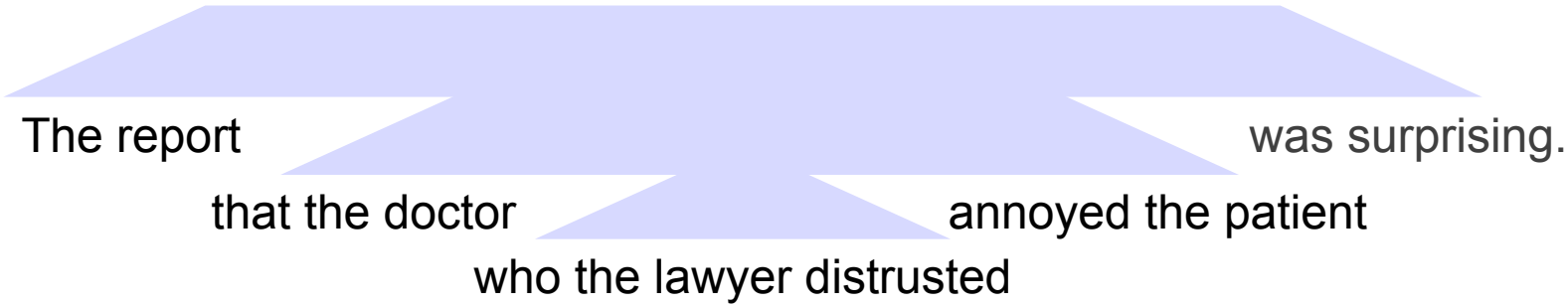
The report was surprising.

easy to comprehend

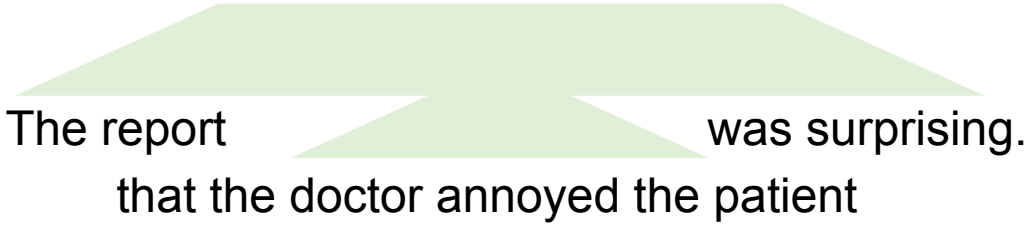
Comprehending Hierarchical Structure

easy to comprehend

Comprehending Hierarchical Structure

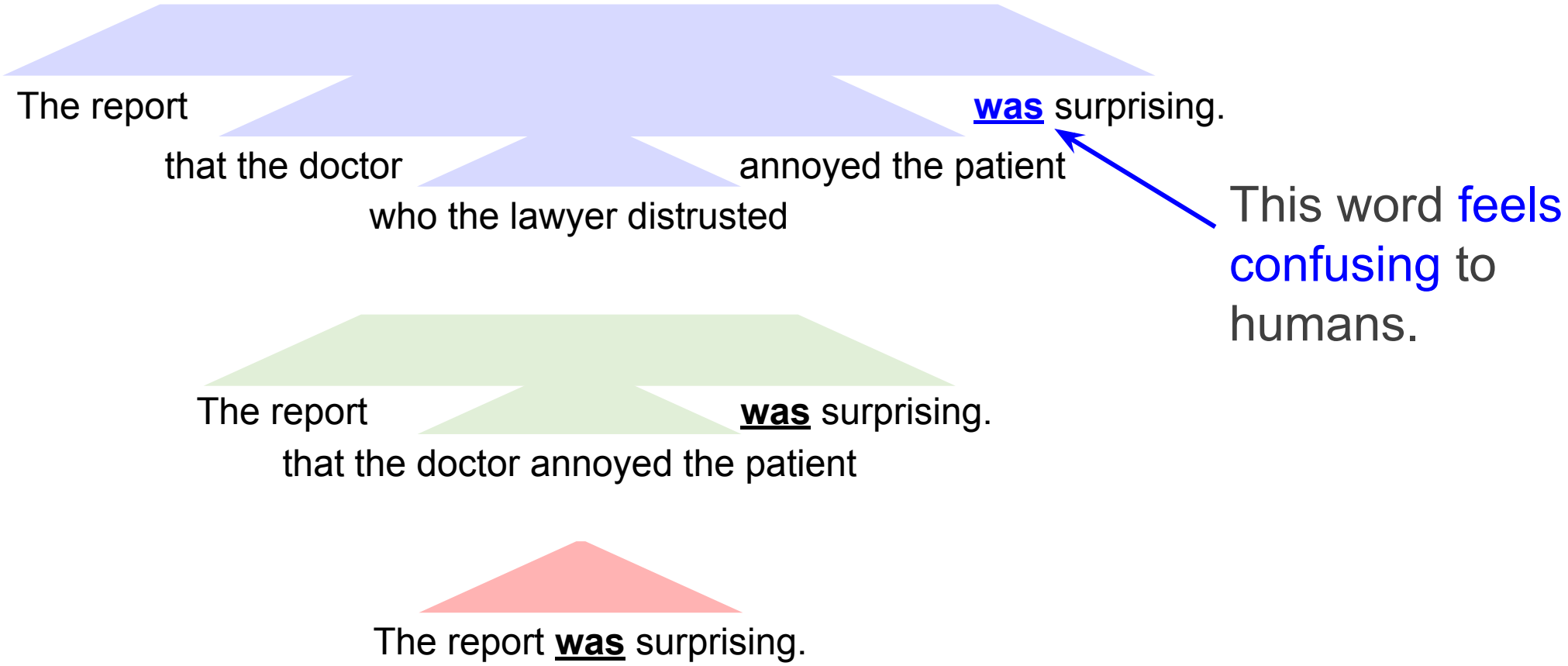


hard to comprehend



easy to comprehend

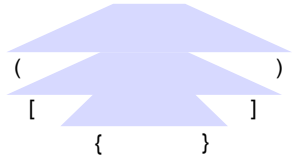
Comprehending Hierarchical Structure



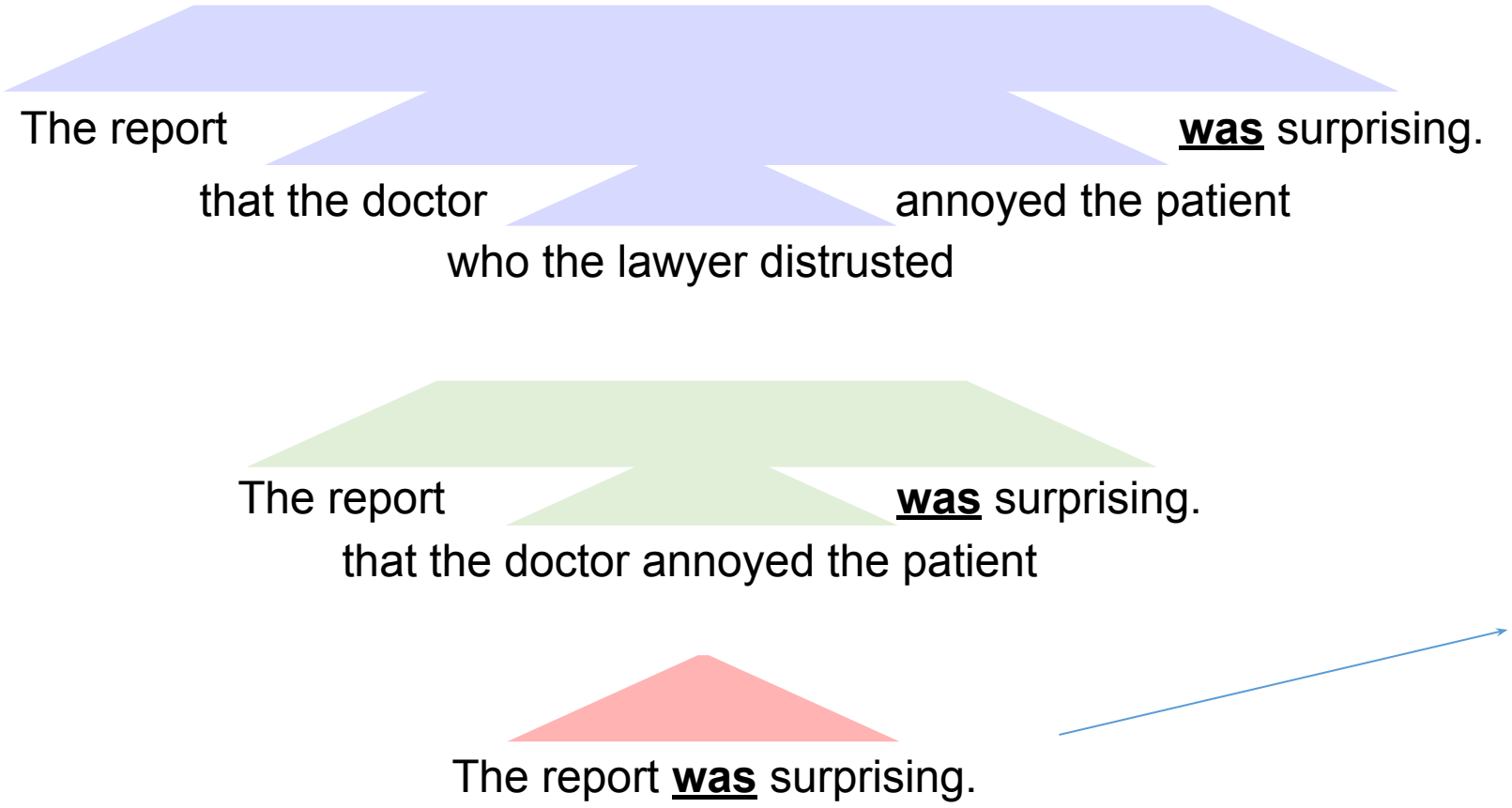
Comprehending Hierarchical Structure

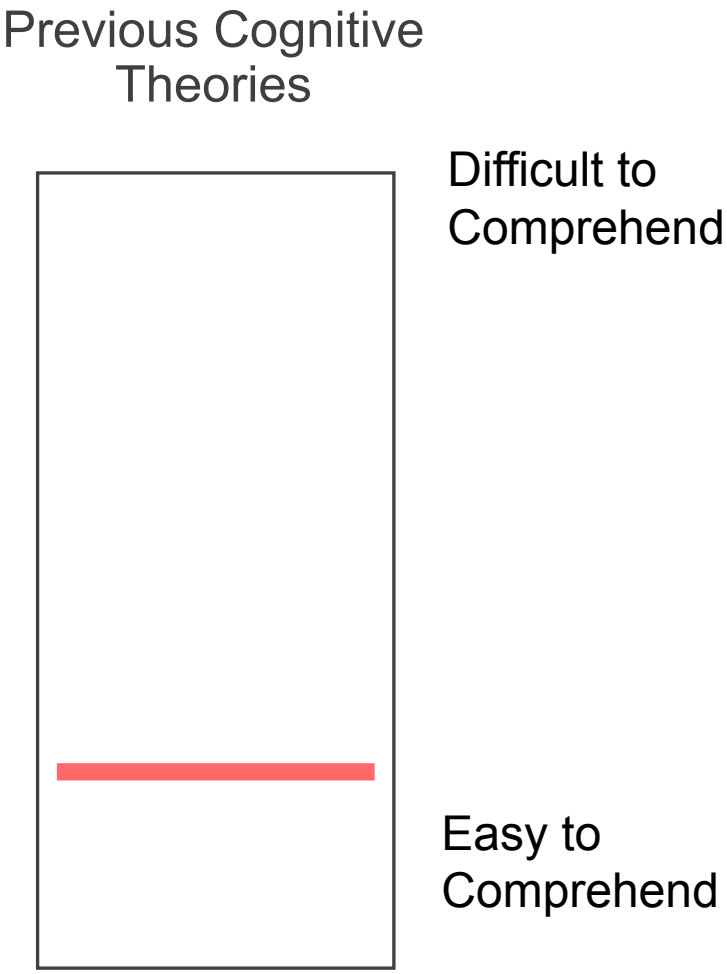
This word feels confusing to humans.

As if humans had trouble counting brackets.



Comprehending Hierarchical Structure



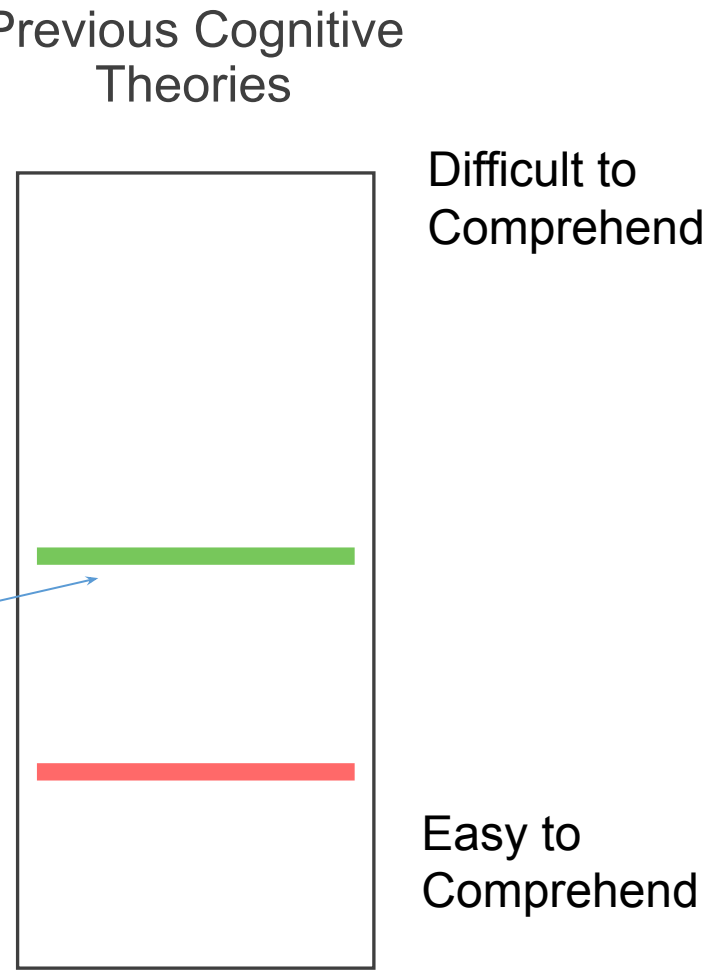


Comprehending Hierarchical Structure

The report that the doctor who the lawyer distrusted annoyed the patient **was** surprising.

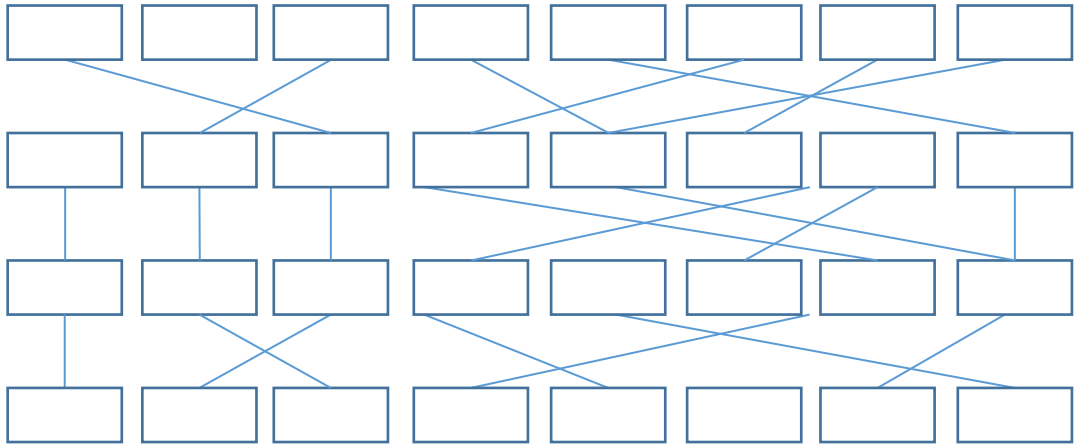
The report that the doctor annoyed the patient **was** surprising.

The report **was** surprising.



Comprehending Hierarchical Structure

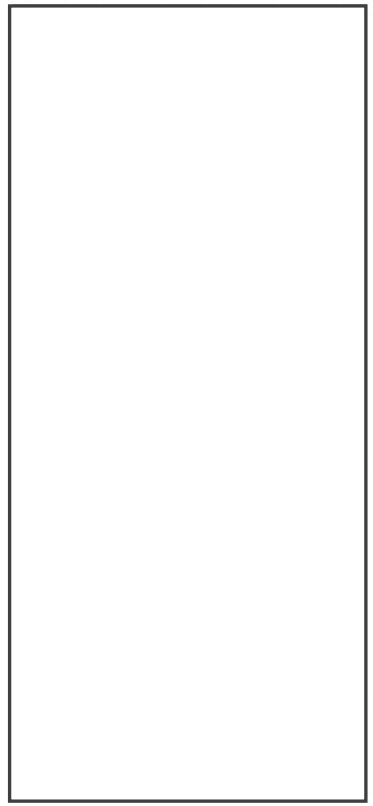
GPT-2 and GPT-3 (Radford et al 2019, Brown et al 2020)



The report that the doctor annoyed the patient was surprising.

Comprehending Hierarchical Structure

GPT-2 and GPT-3



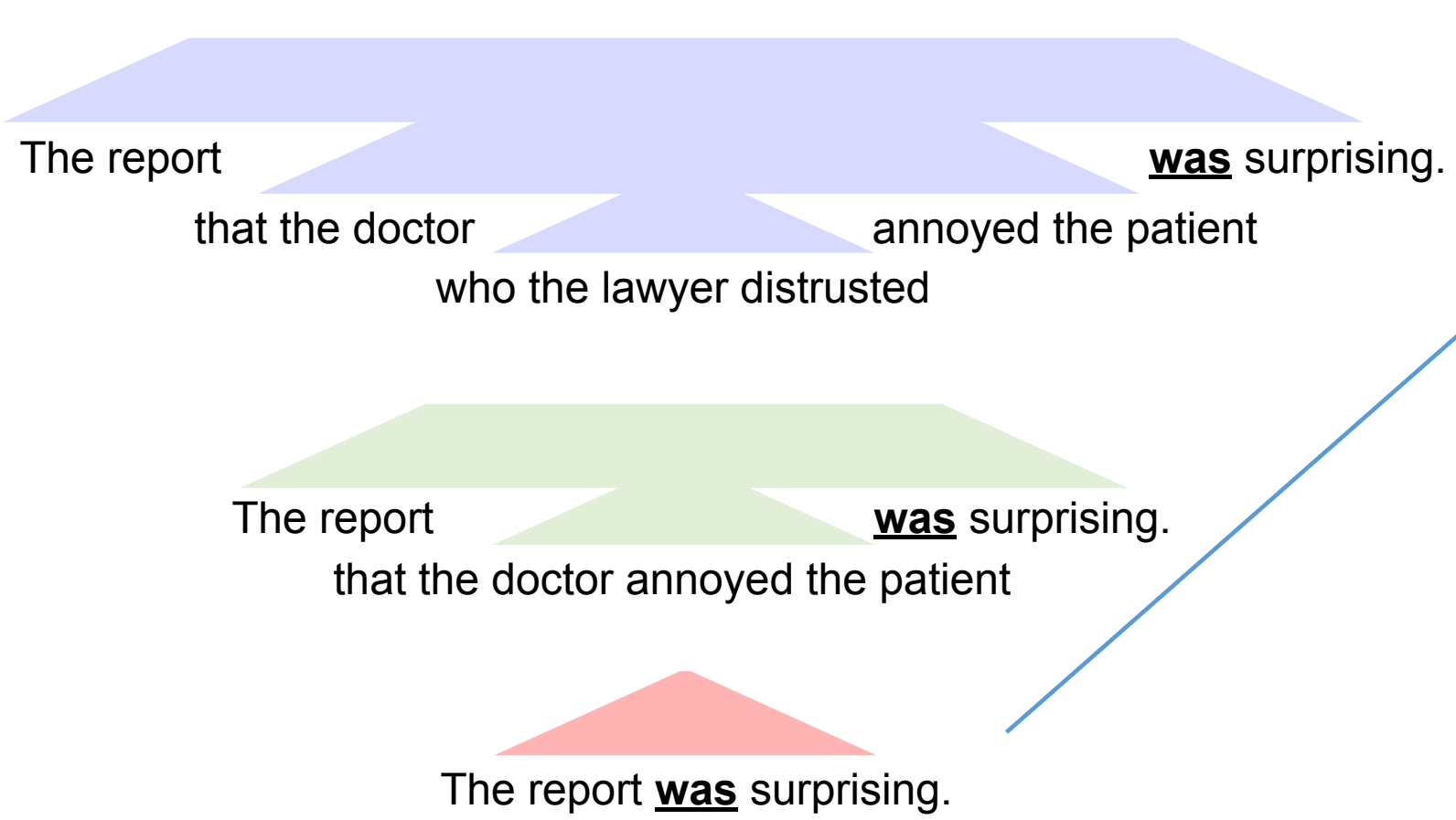
Surprising

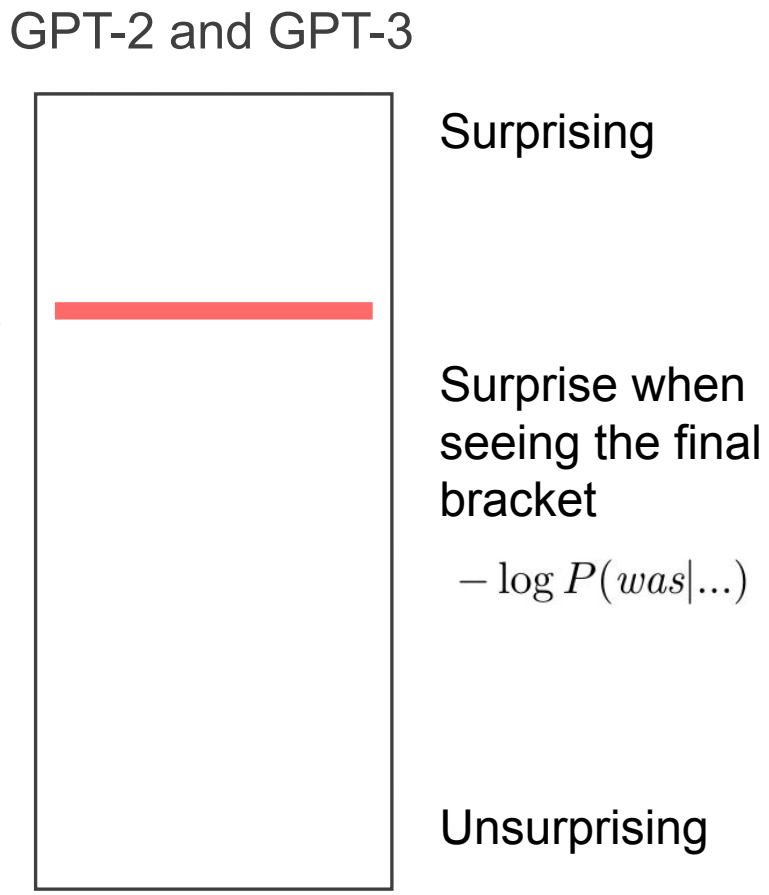
Surprise when seeing the final bracket

$-\log P(was|...)$

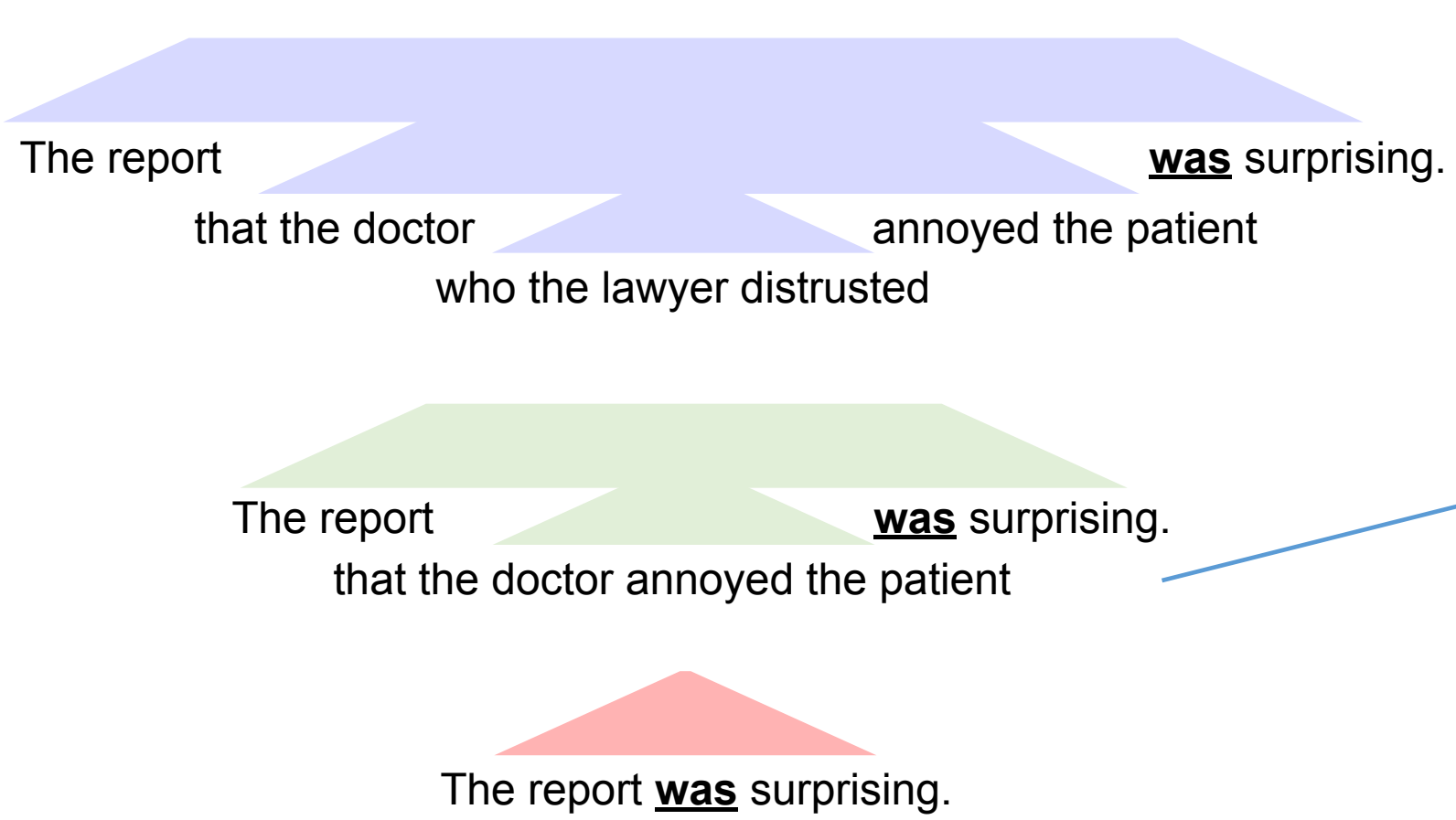
Unsurprising

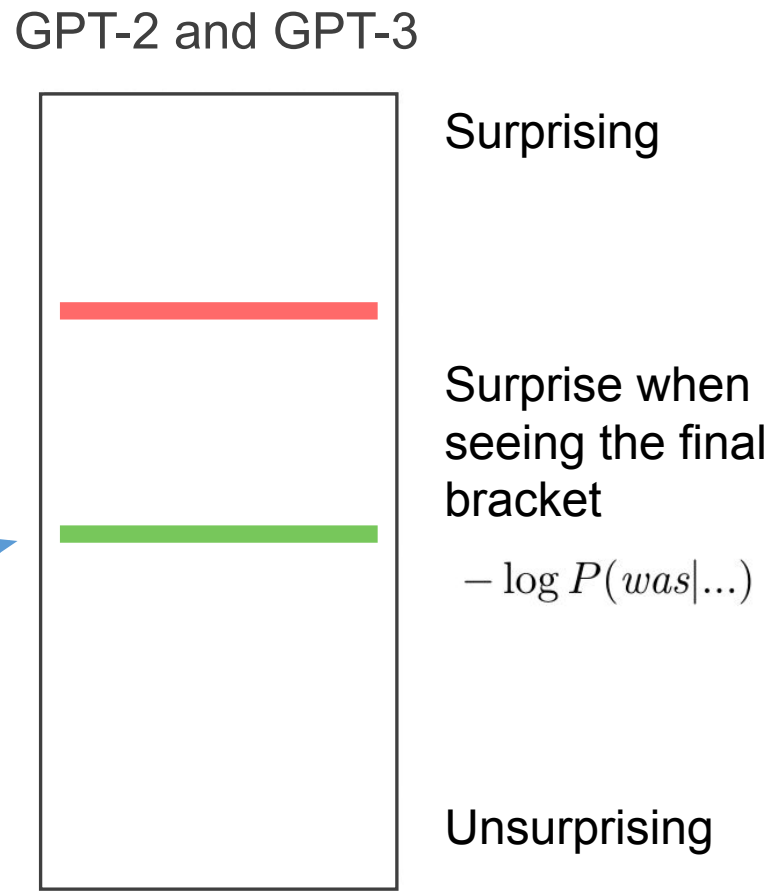
Comprehending Hierarchical Structure



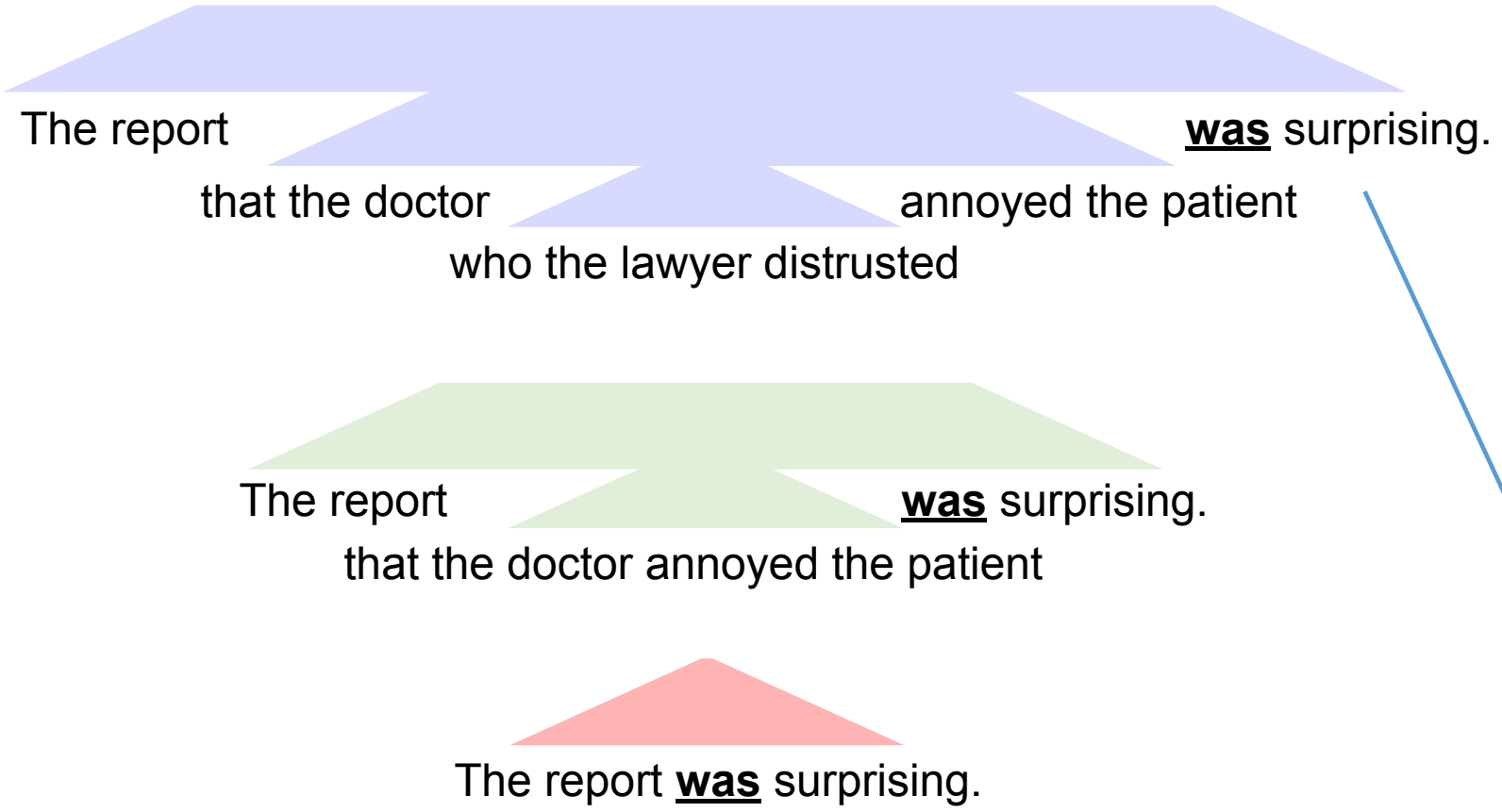


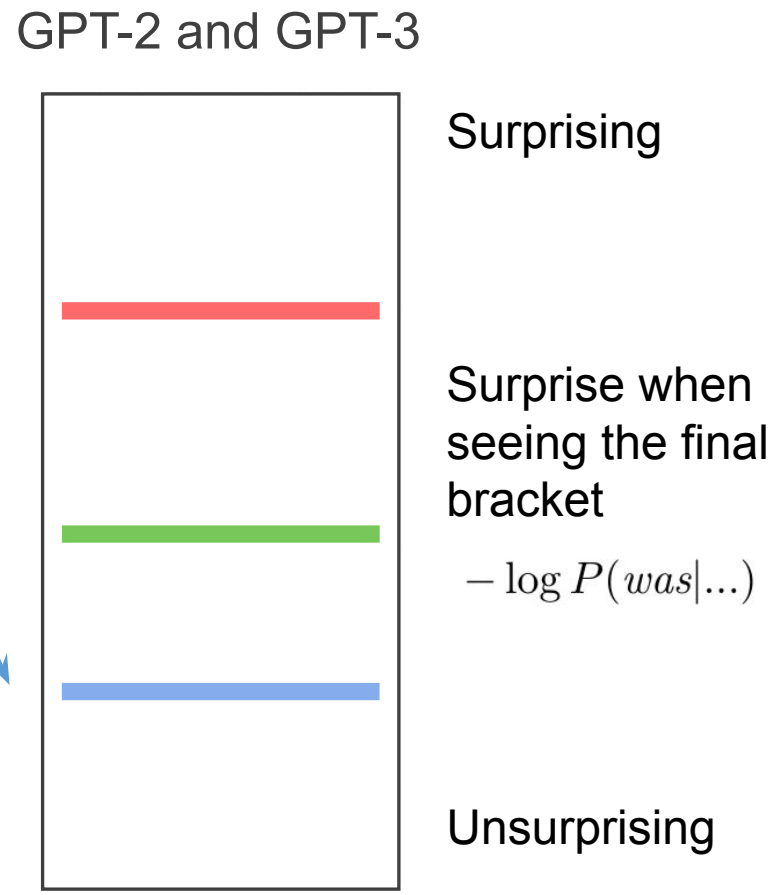
Comprehending Hierarchical Structure



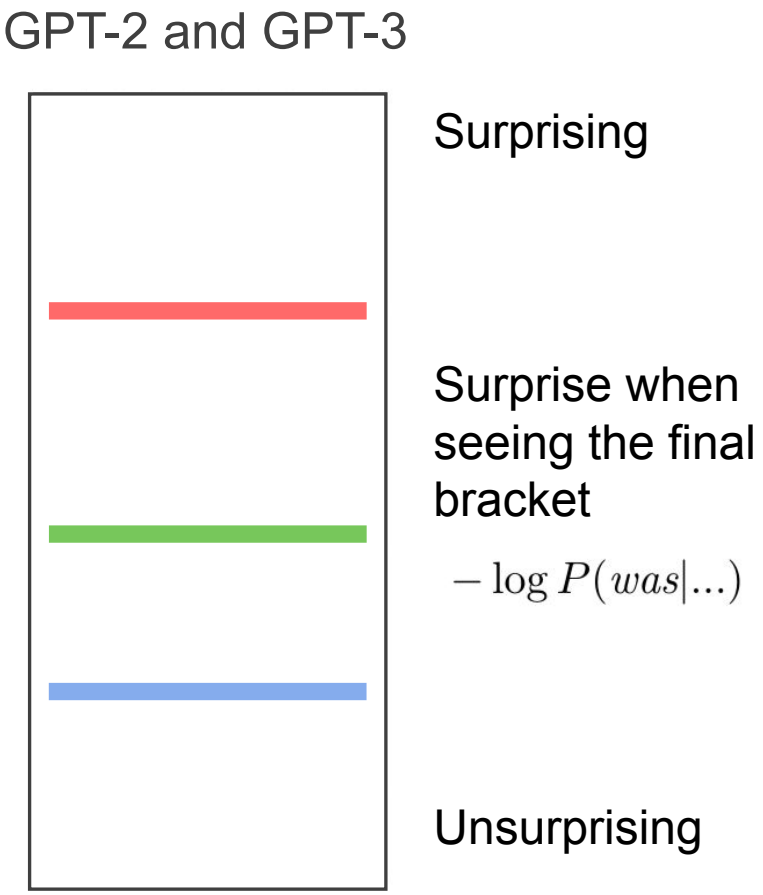


Comprehending Hierarchical Structure



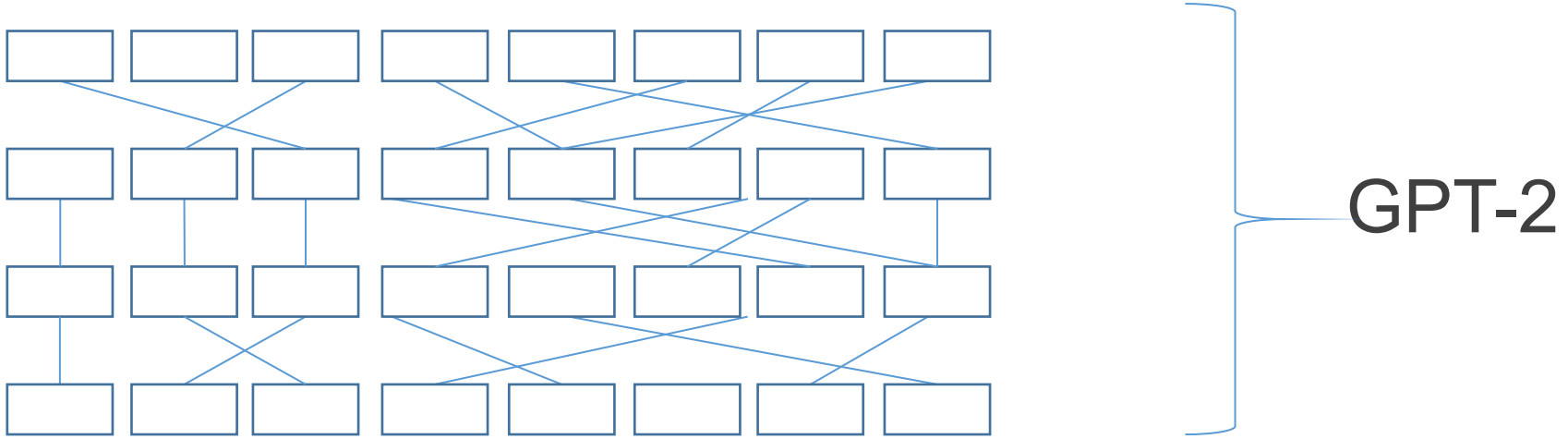


Comprehending Hierarchical Structure



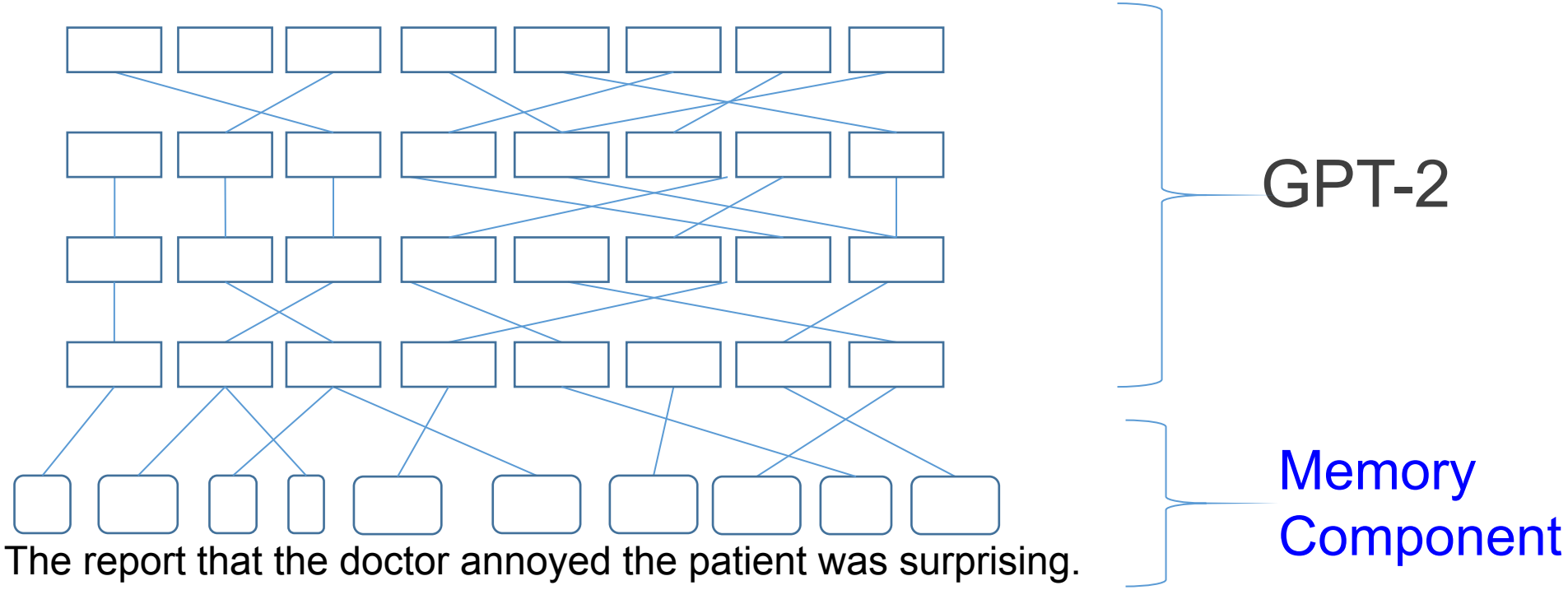
Not human-like!

Memory-Constrained GPT-2

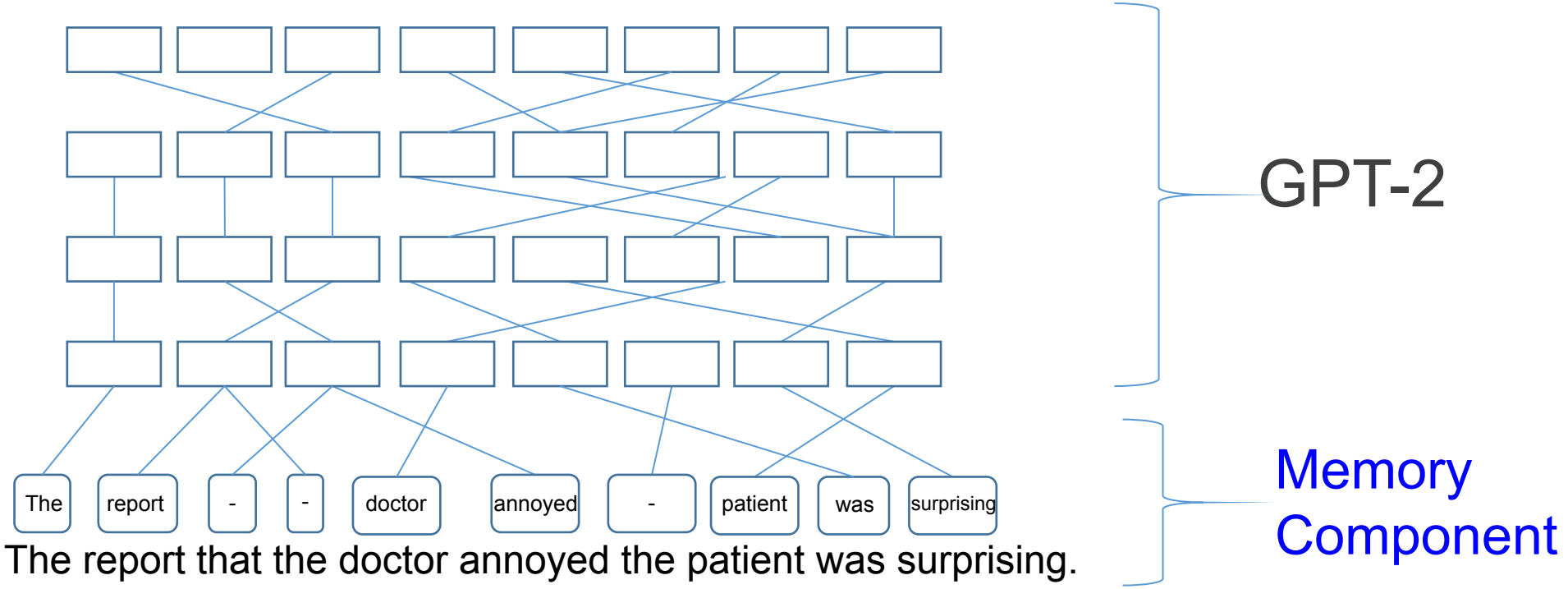


The report that the doctor annoyed the patient was surprising.

Memory-Constrained GPT-2



Memory-Constrained GPT-2



The report that the doctor annoyed the patient was surprising.

Memory-Constrained GPT-2

Training Objective: Memory-Constrained GPT-2

$$\arg\min \mathbb{E}\left[- \log P(\text{input}) + \alpha \, \#\text{RememberedWords} \right]$$

Memory-Constrained GPT-2

Training Objective: Memory-Constrained GPT-2

$$\arg\min \mathbb{E}\left[- \log P(\text{input}) + \alpha \, \#\text{RememberedWords} \right]$$

log-likelihood of input text

Memory-Constrained GPT-2

Training Objective: Memory-Constrained GPT-2

$$\arg\min \mathbb{E}\left[- \log P(\text{input}) + \alpha \, \#\text{RememberedWords} \right]$$

log-likelihood of input text

memory load

Memory-Constrained GPT-2

Training Objective: Memory-Constrained GPT-2

$$\arg\min \mathbb{E}\left[- \log P(\text{input}) + \alpha \, \#\text{RememberedWords} \right]$$

Trained using reinforcement learning (policy gradient).

Memory-Constrained GPT-2

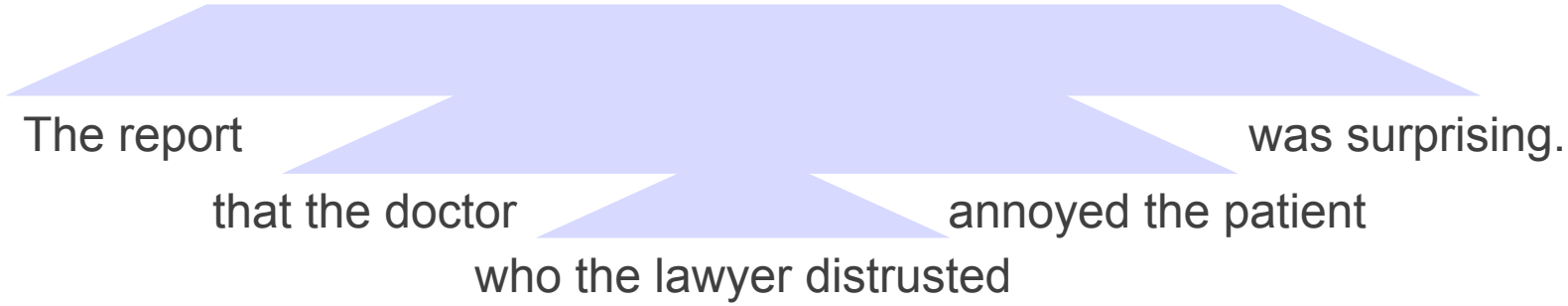
Training Objective: Memory-Constrained GPT-2

$$\arg\min \mathbb{E}\left[- \log P(\text{input}) + \alpha \, \#\text{RememberedWords} \right]$$

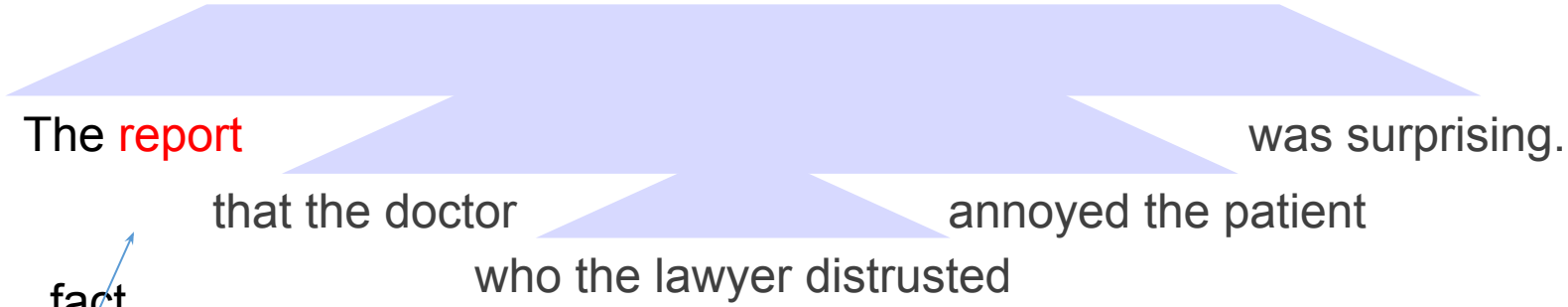
Trained using reinforcement learning (policy gradient).

Unsupervised: No human data needed.

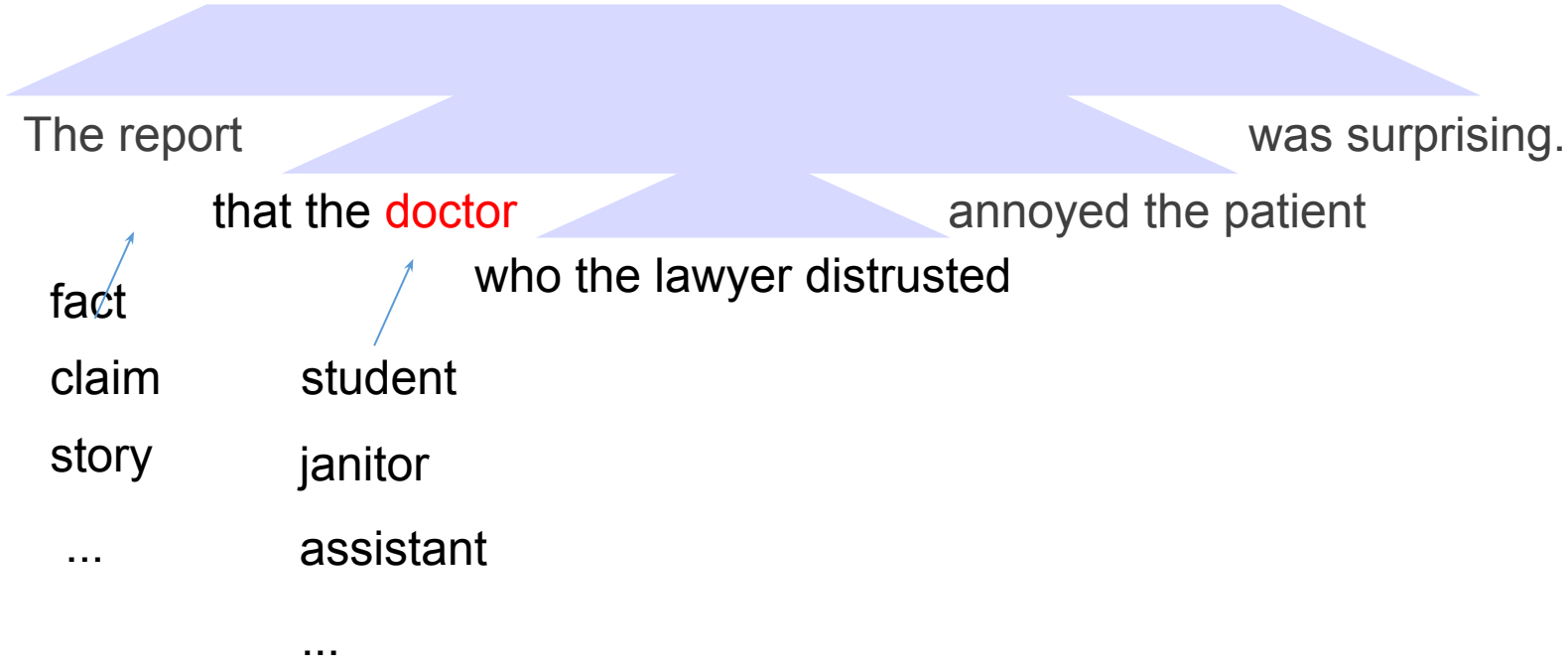
Stimulus Set



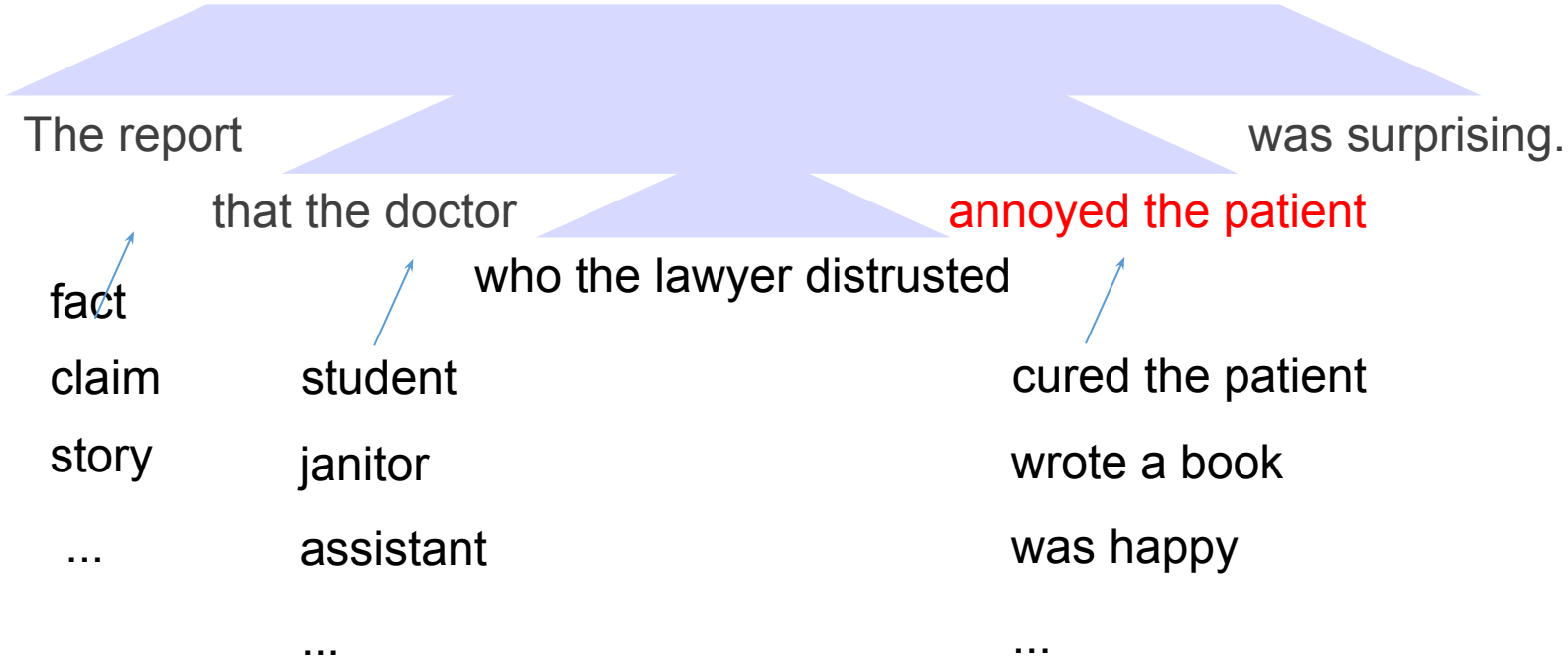
Stimulus Set



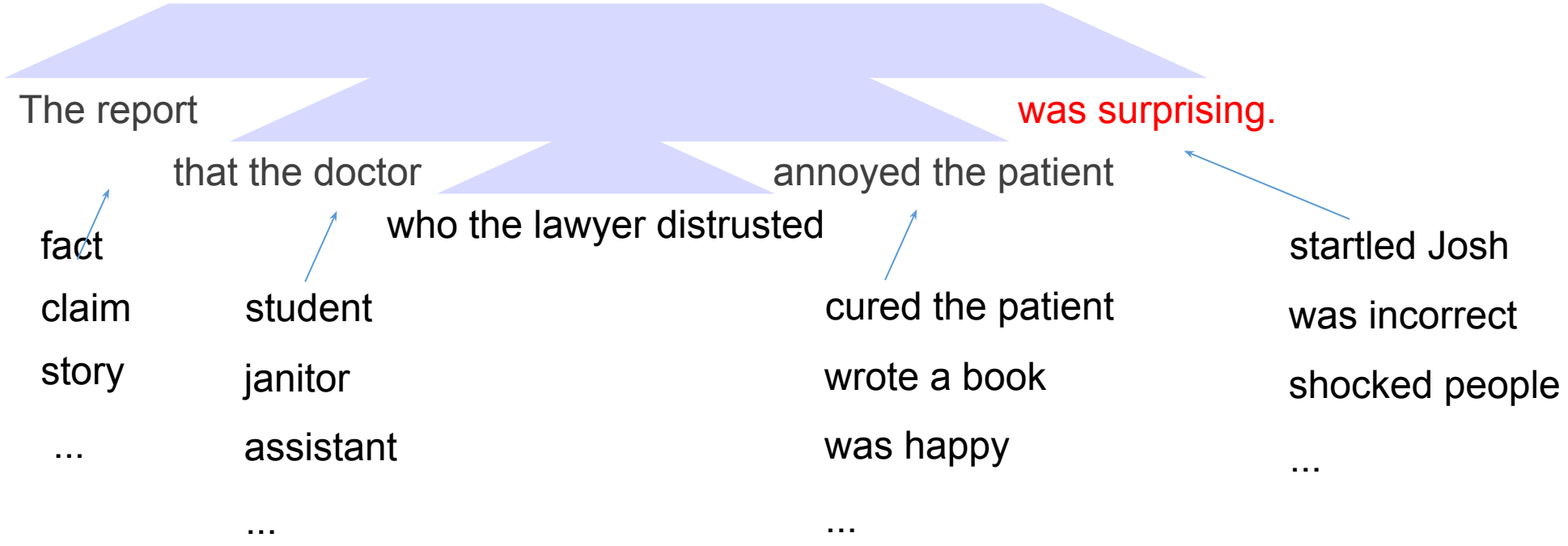
Stimulus Set



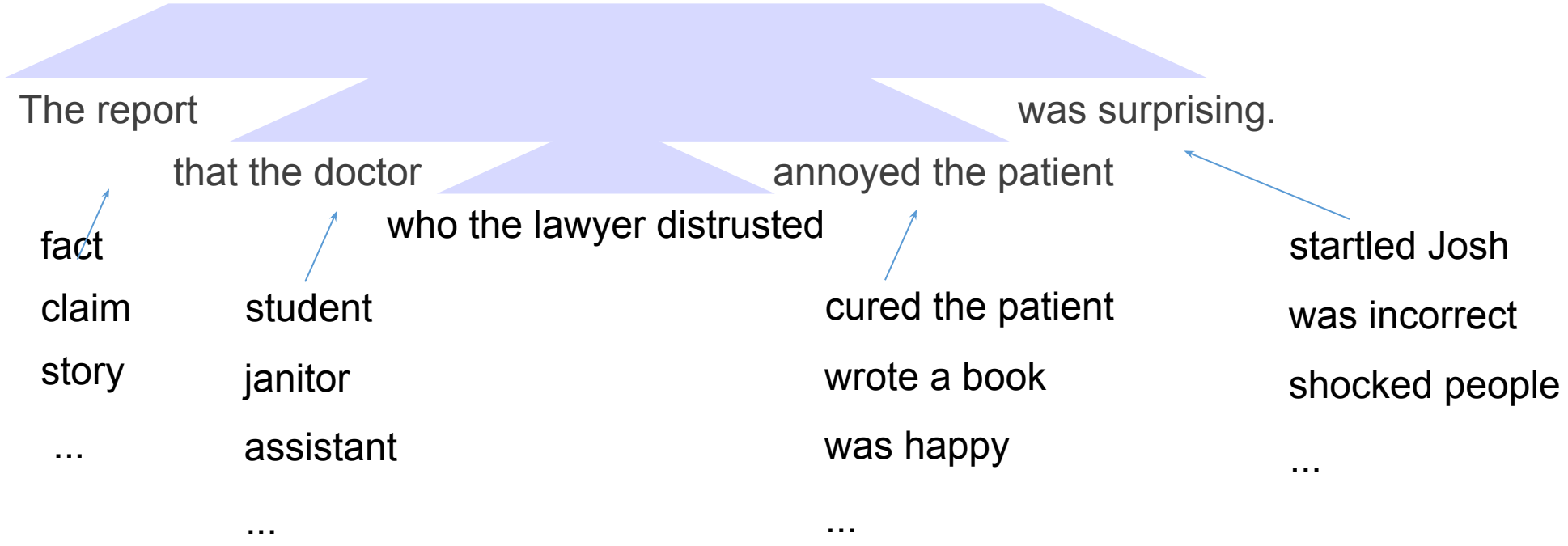
Stimulus Set



Stimulus Set

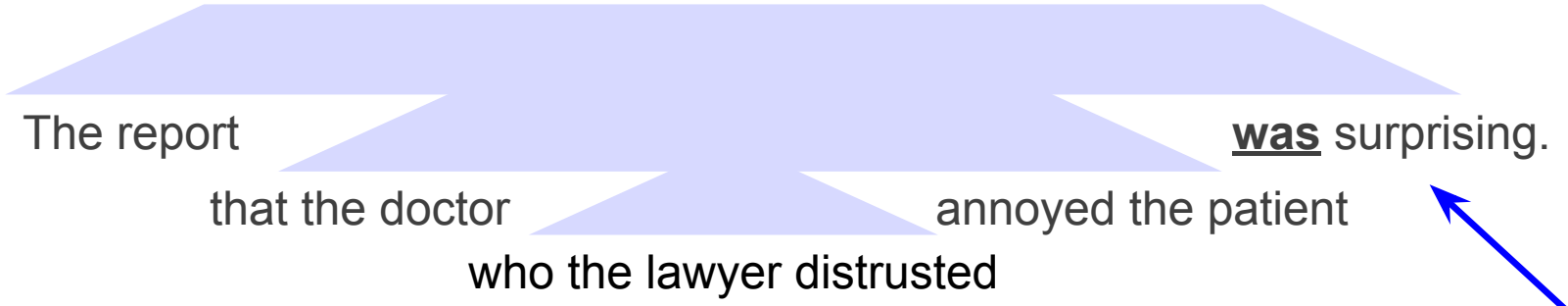


Stimulus Set



Total of ~10K different sentences

Stimulus Set



How long does it take a human to process this part?

Stimulus Set

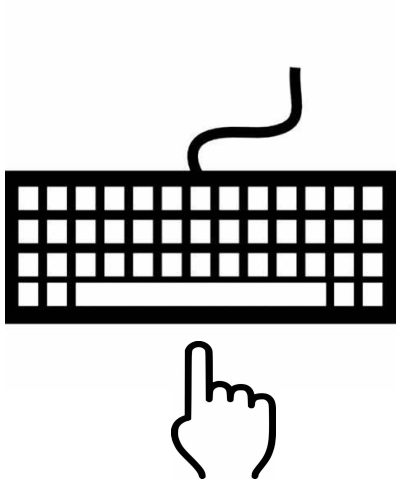
The report
that the doctor
who the lawyer distrusted
annoyed the patient
was surprising.

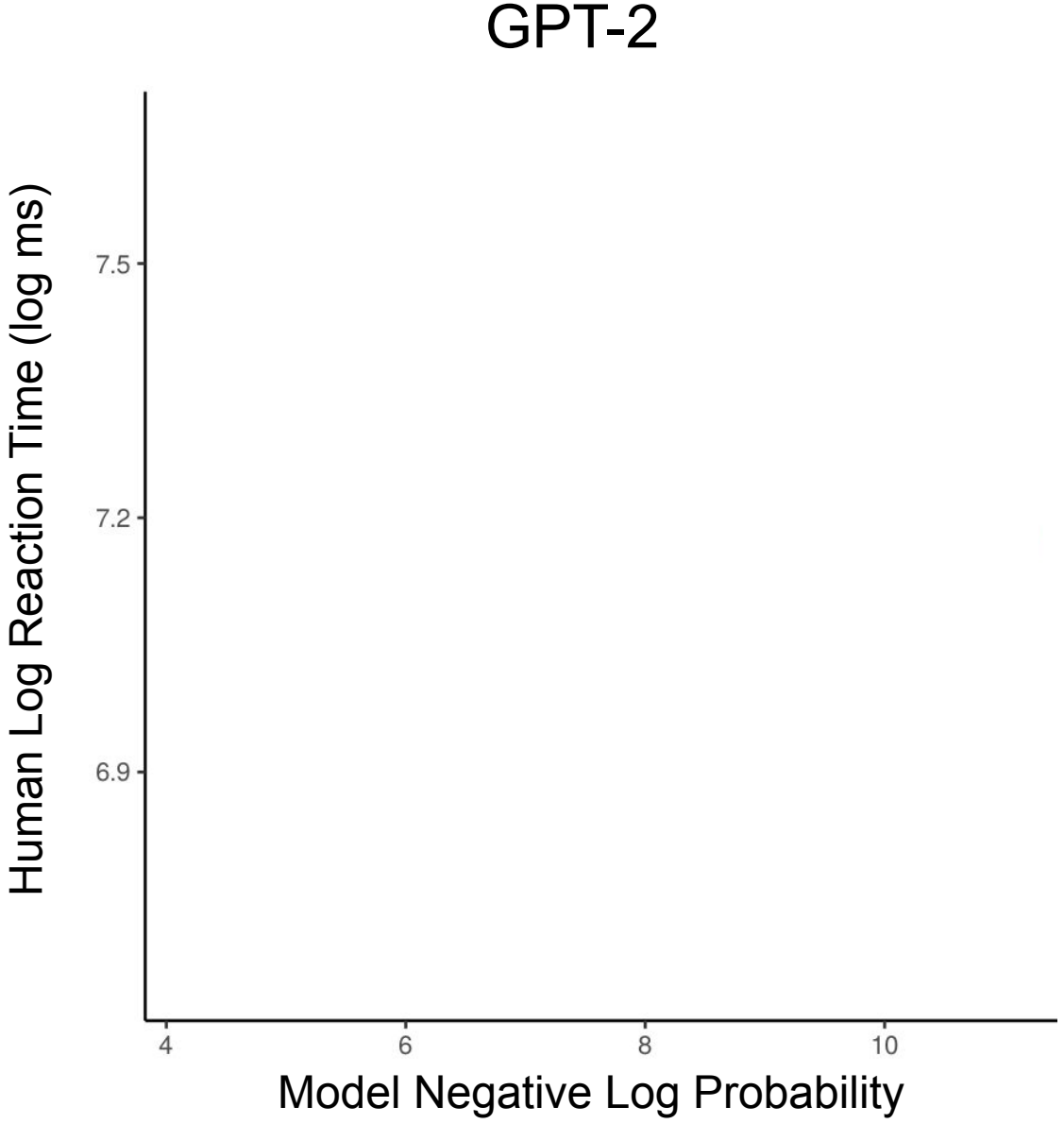
How long does it take a human to process this part?

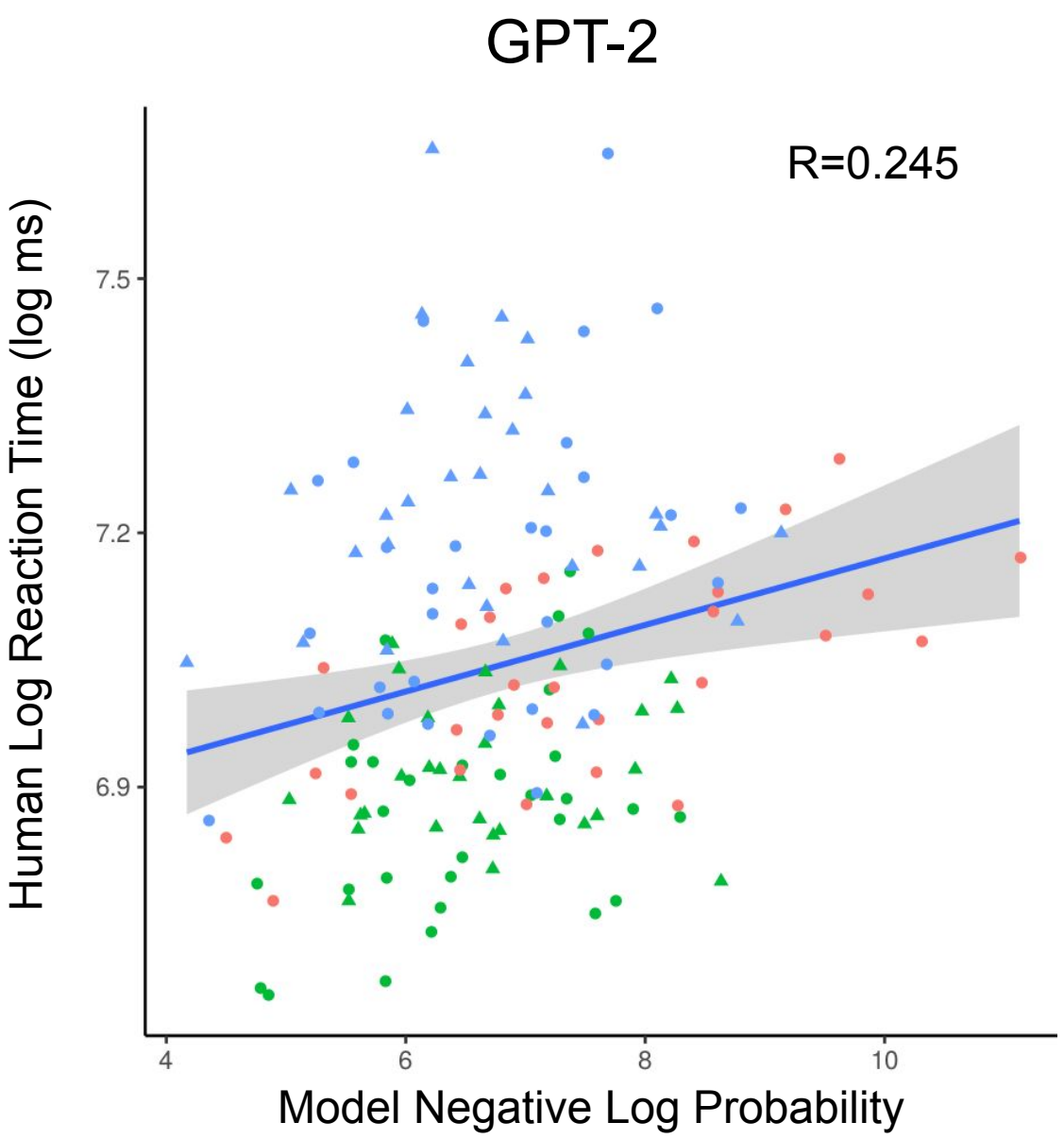
Crowd-sourced experiment

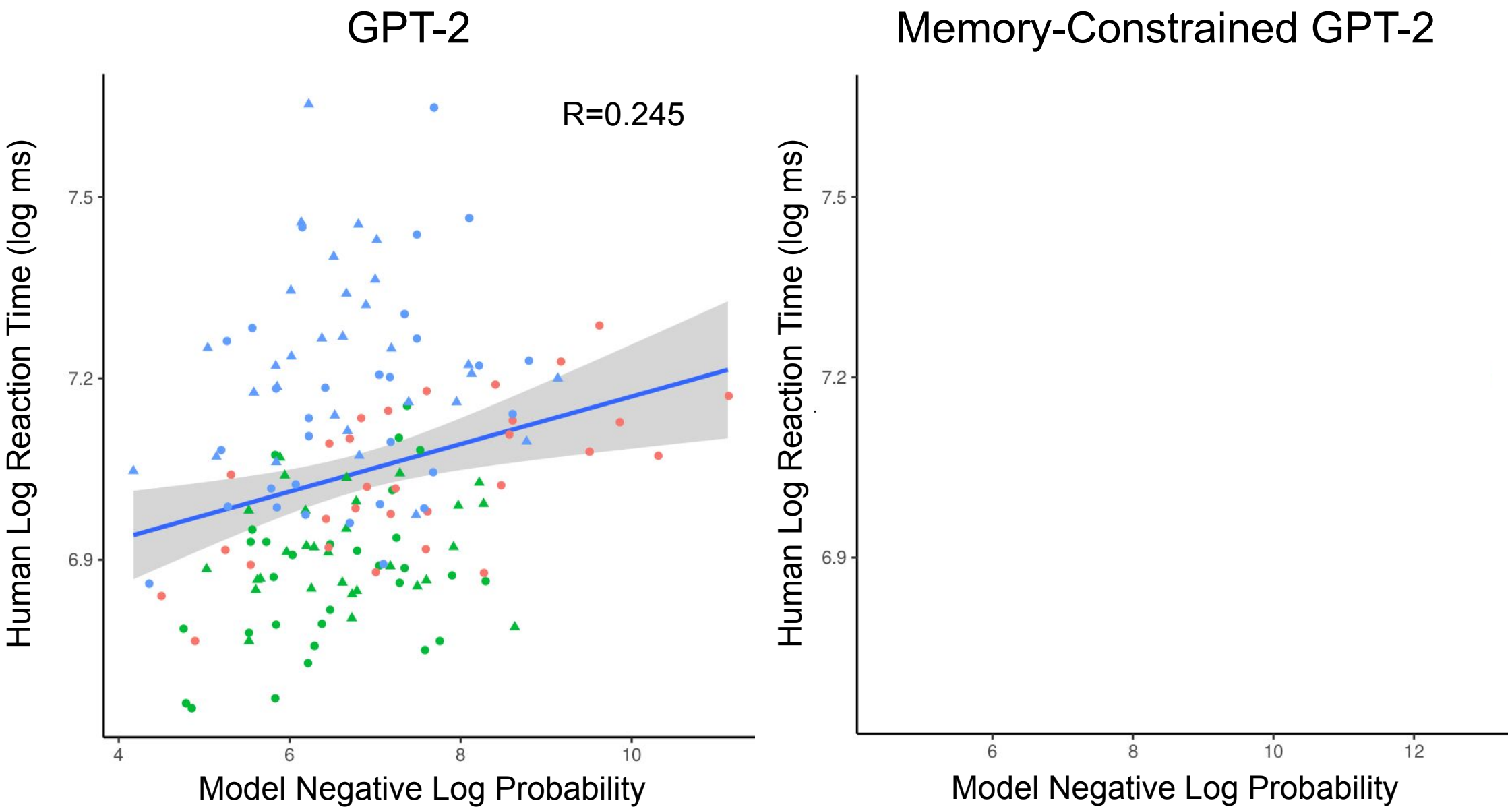
1000 participants

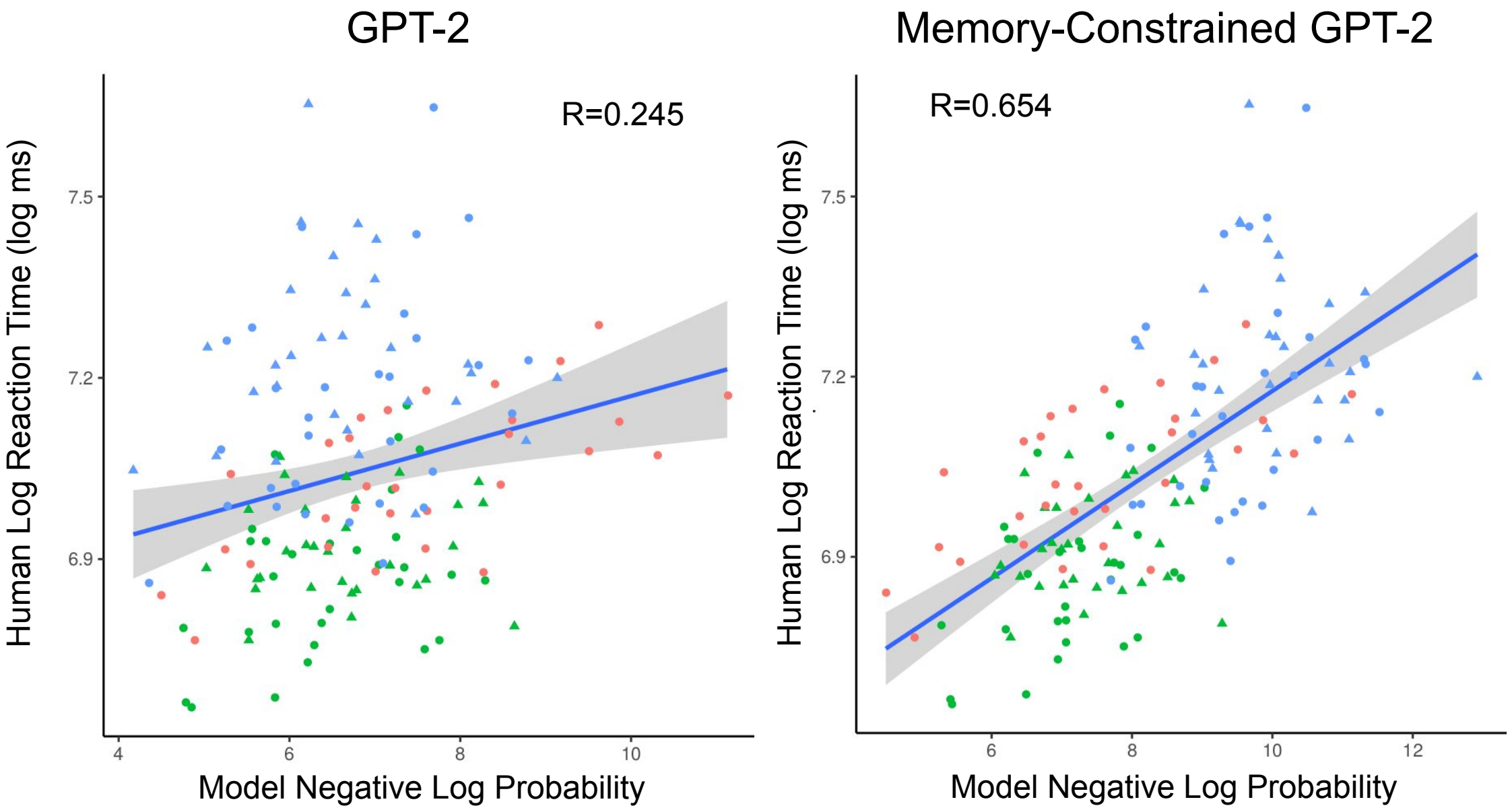
Each read 10 sentences




GPT-2
Human Log Reaction Time (log ms)
7.5
7.2
6.9
4
6
8
10
Model Negative Log Probability


GPT-2
R=0.245
Human Log Reaction Time (log ms)
7.5
7.2
6.9
4
6
8
10
Model Negative Log Probability

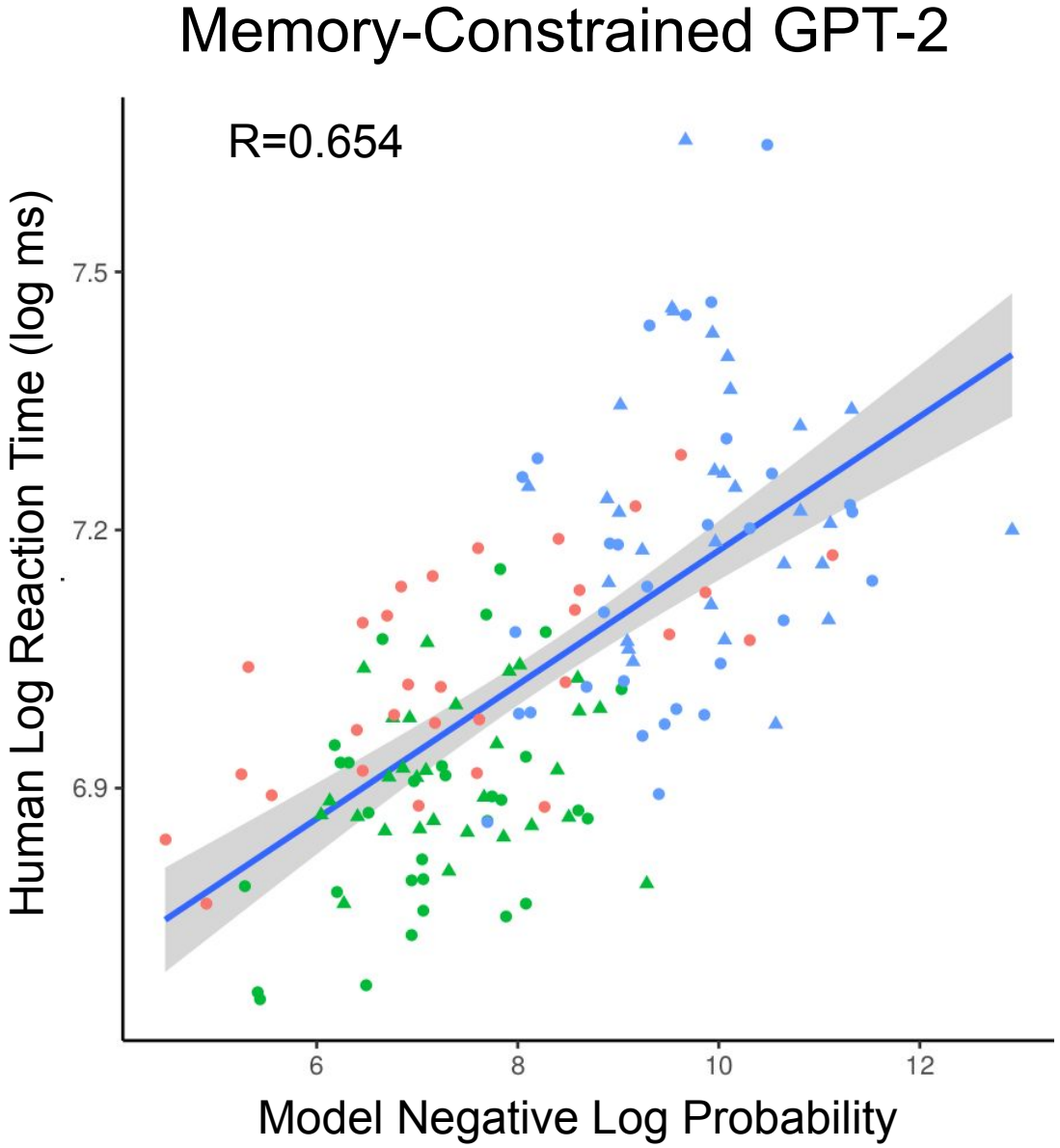

GPT-2
R=0.245
Human Log Reaction Time (log ms)
7.5
7.2
6.9
4
6
8
10
Model Negative Log Probability
Memory-Constrained GPT-2
Human Log Reaction Time (log ms)
7.5
7.2
6.9
6
8
10
12
Model Negative Log Probability


GPT-2
R=0.245
Human Log Reaction Time (log ms)
7.5
7.2
6.9
4
6
8
10
Model Negative Log Probability
Memory-Constrained GPT-2
R=0.654
Human Log Reaction Time (log ms)
7.5
7.2
6.9
6
8
10
12
Model Negative Log Probability

Unsupervised human-like memory component

Correctly predicts when humans find hierarchical structure difficult

Outperforms GPT-2 and cognitive theories



Languages vary a lot in how they order information

Languages vary a lot in how they order information

katabt	risāla	li	sadīq
VERB	NOUN	ADP	NOUN
wrote	letter	to	friend

Languages vary a lot in how they order information

katabt	risāla	li	sadīq
VERB	NOUN	ADP	NOUN
wrote	letter	to	friend

tomodachi	ni	tegami-o	kaita
NOUN	ADP	NOUN	VERB
friend	to	letter	wrote

Languages vary a lot in how they order information

katabt	risāla	li	sadīq
VERB	NOUN	ADP	NOUN
wrote	letter	to	friend

tomodachi	ni	tegami-o	kaita
NOUN	ADP	NOUN	VERB
friend	to	letter	wrote

but not without bounds.

Correlation Universals

Arabic	katabt	risāla	li	ṣadīq
	wrote	letter	to	friend
Japanese	tomodachi	ni	tegami-o	kaita
	friend	to	letter	wrote

(Dryer 1993, 2013)

Correlation Universals

Object-Verb

letter wrote

Verb-Object

wrote letter

Arabic	katabt	risāla	li	sadīq
	wrote	letter	to	friend
Japanese	tomodachi	ni	tegami-o	kaita
	friend	to	letter	wrote

(Dryer 1993, 2013)

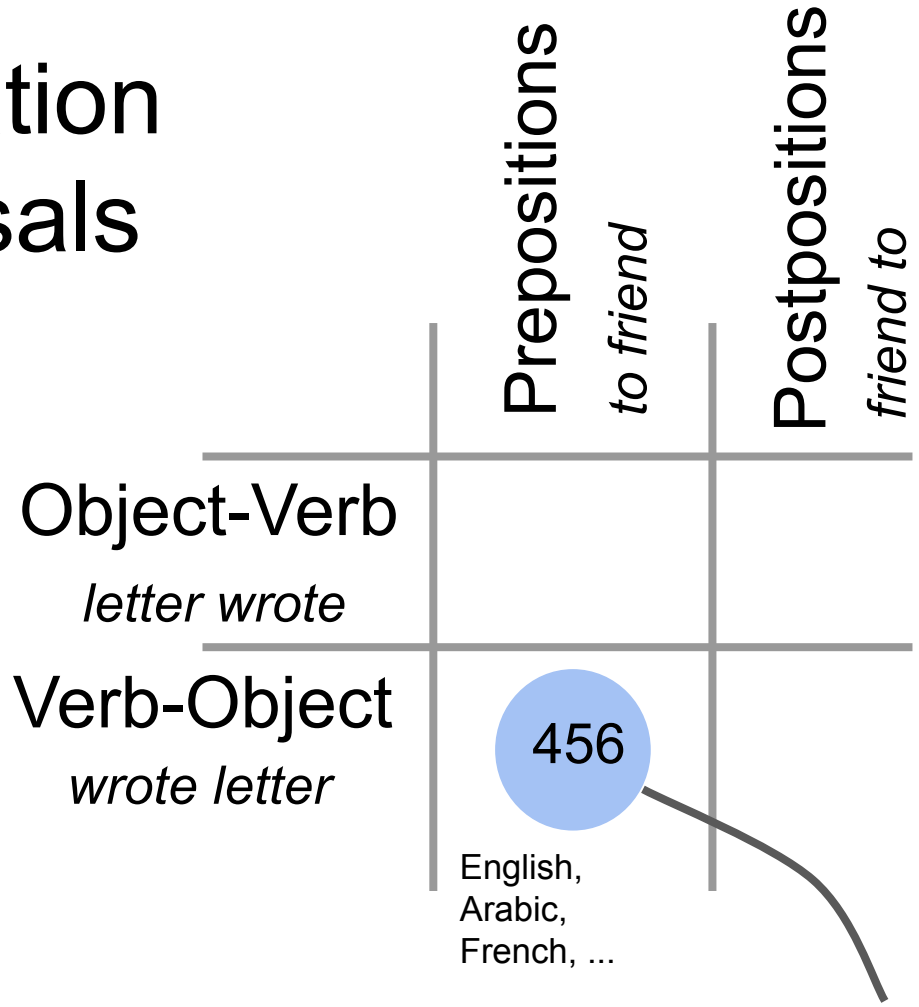
Correlation Universals

	Prepositions *to friend*	Postpositions *friend to*
Object-Verb *letter wrote*		
Verb-Object *wrote letter*		

Arabic	katabt	risāla	li	sadīq
	wrote	letter	to	friend
Japanese	tomodachi	ni	tegami-o	kaita
	friend	to	letter	wrote

(Dryer 1993, 2013)

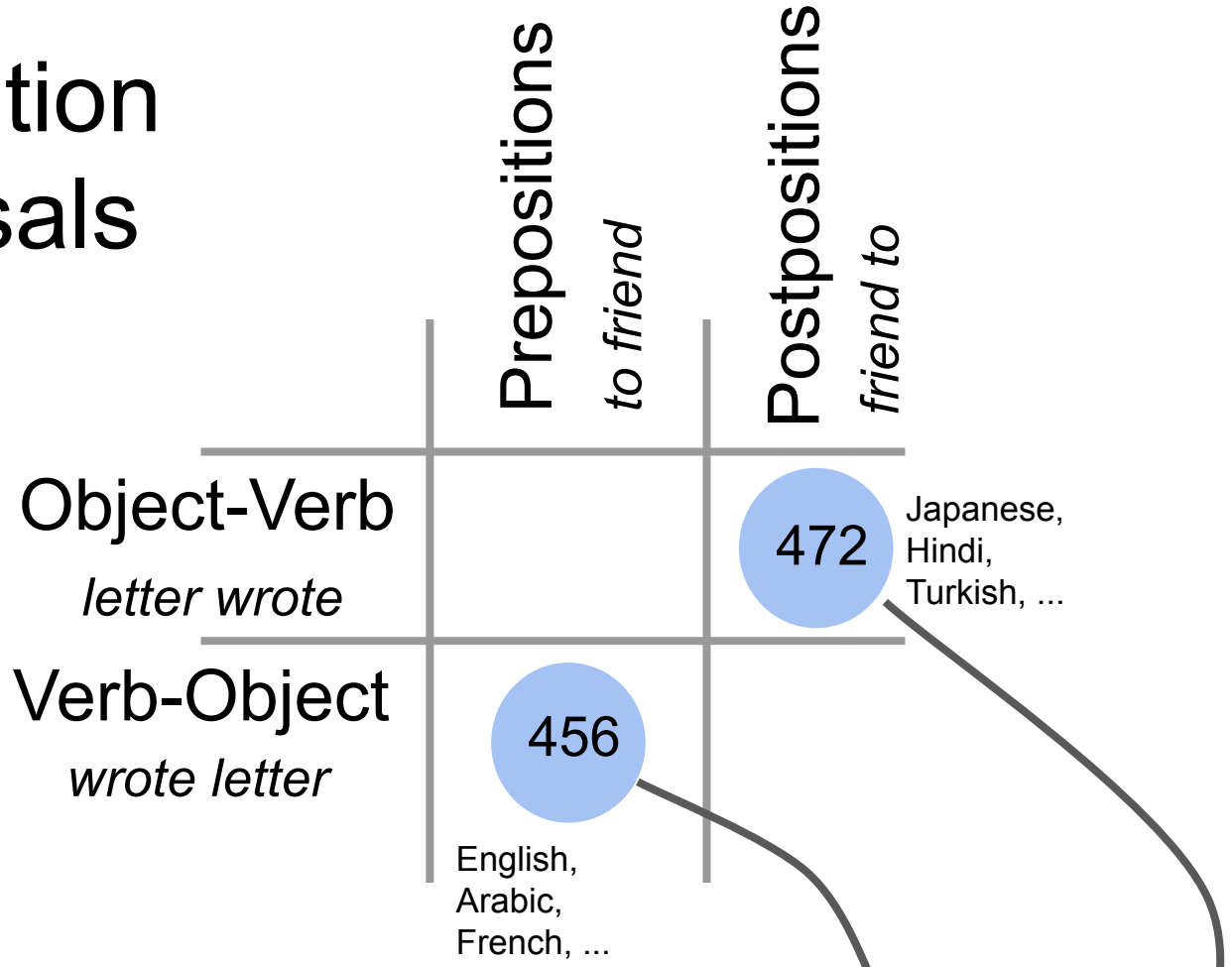
Correlation Universals



Arabic	katabt wrote	risāla letter	li to	sadīq friend
Japanese	tomodachi friend	ni to	tegami-o letter	kaita wrote

(Dryer 1993, 2013)

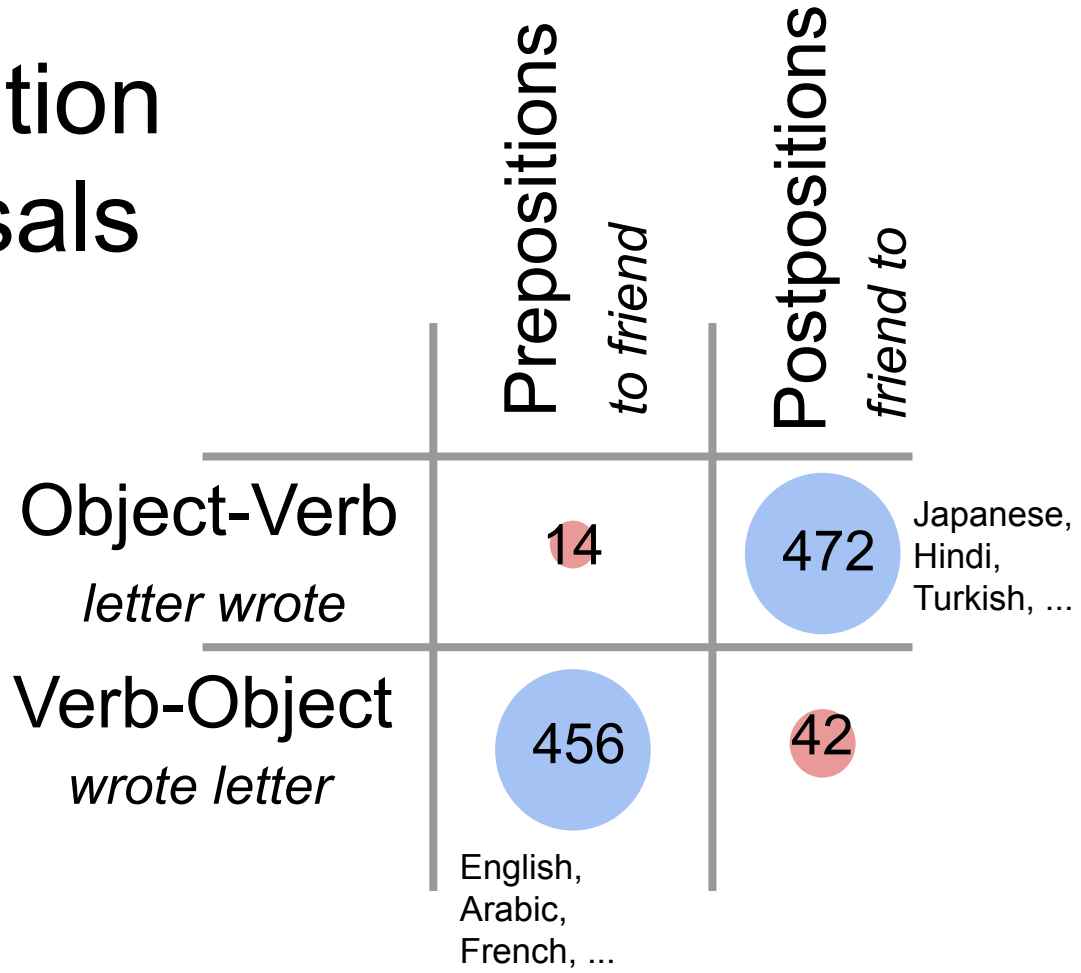
Correlation Universals



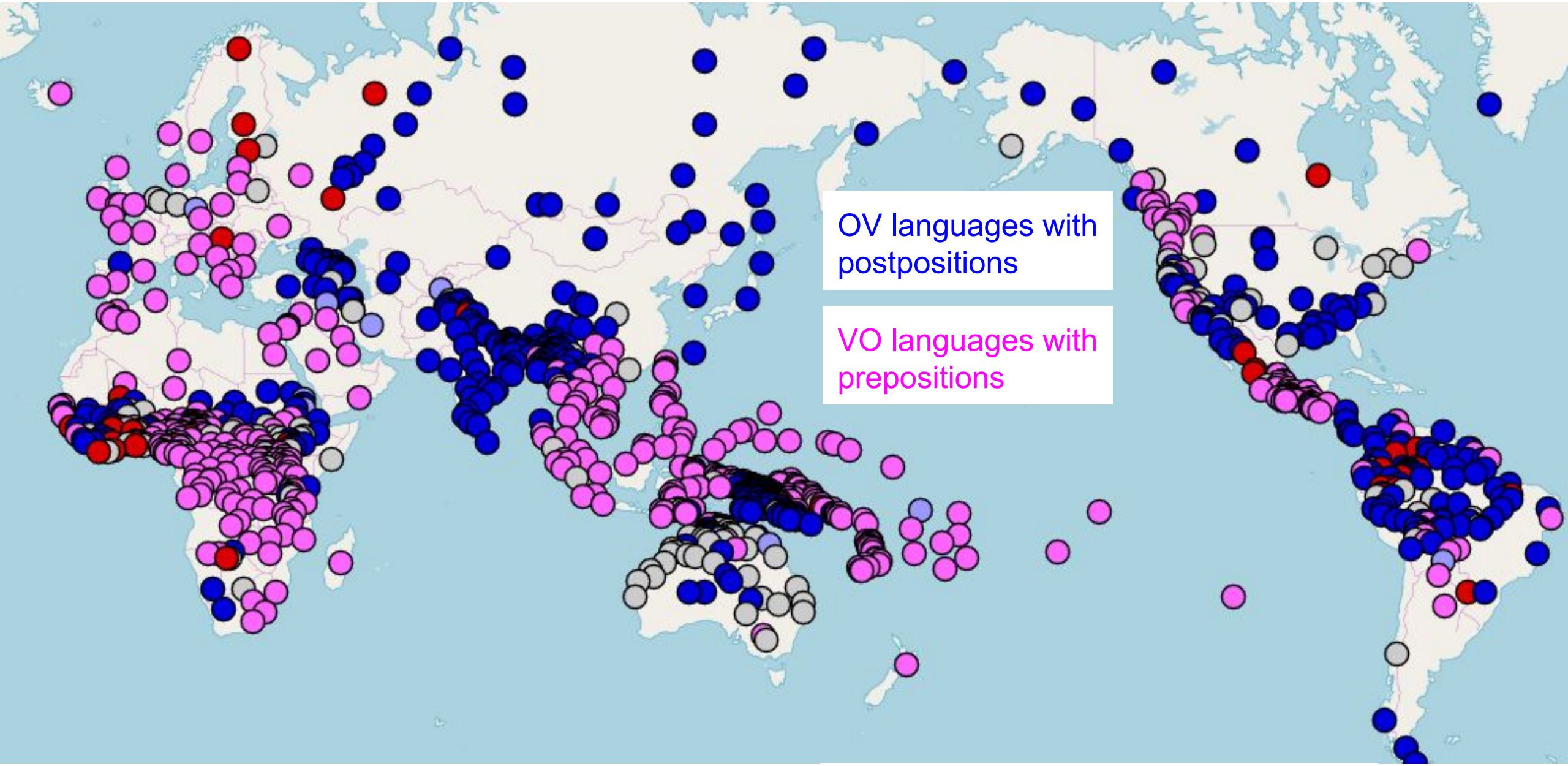
Arabic	katabt	risāla	li	sadīq	
	wrote	letter	to	friend	
Japanese	tomodachi	ni	tegami-o	kaita	
	friend	to	letter	wrote	

(Dryer 1993, 2013)

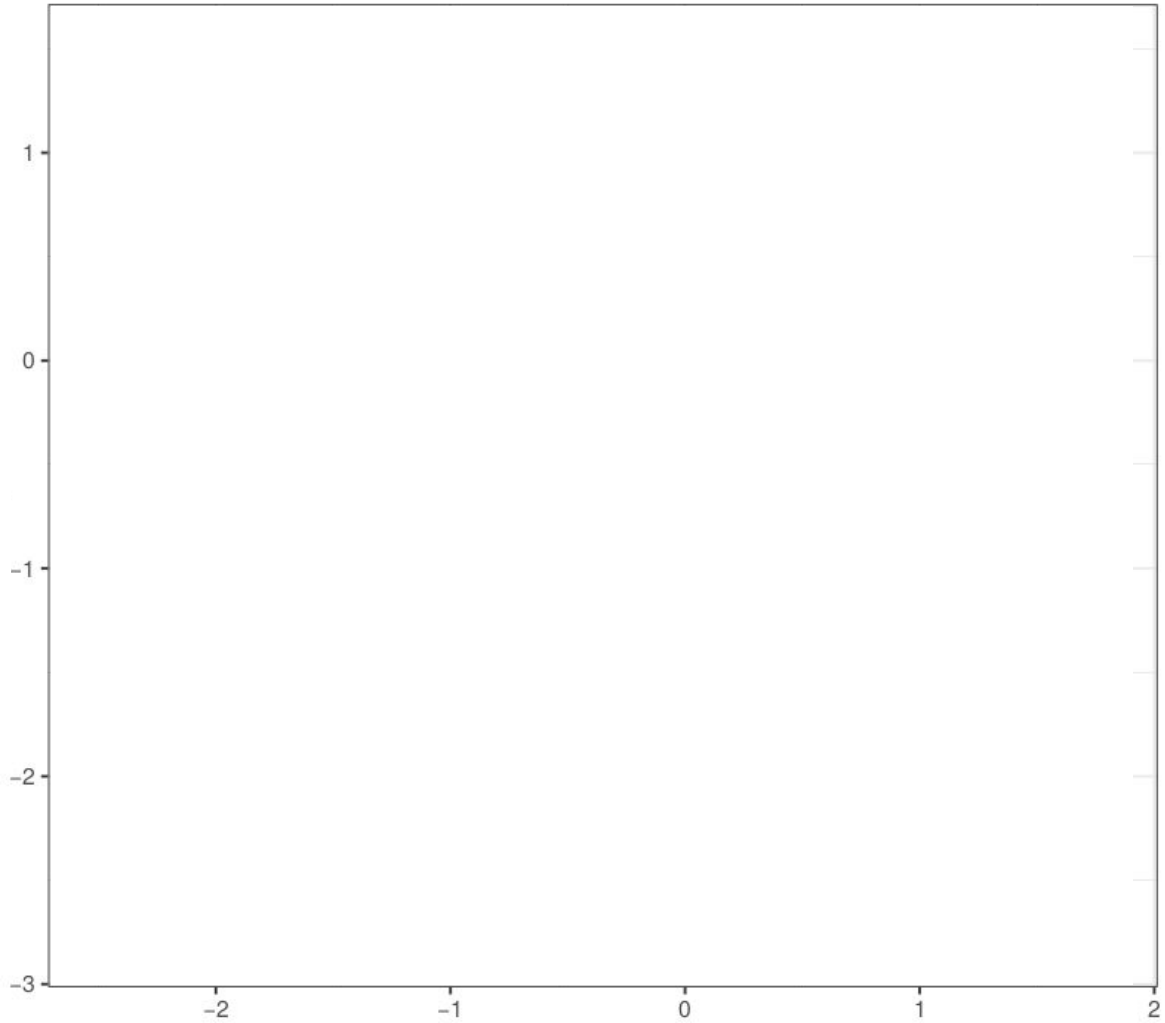
Correlation Universals

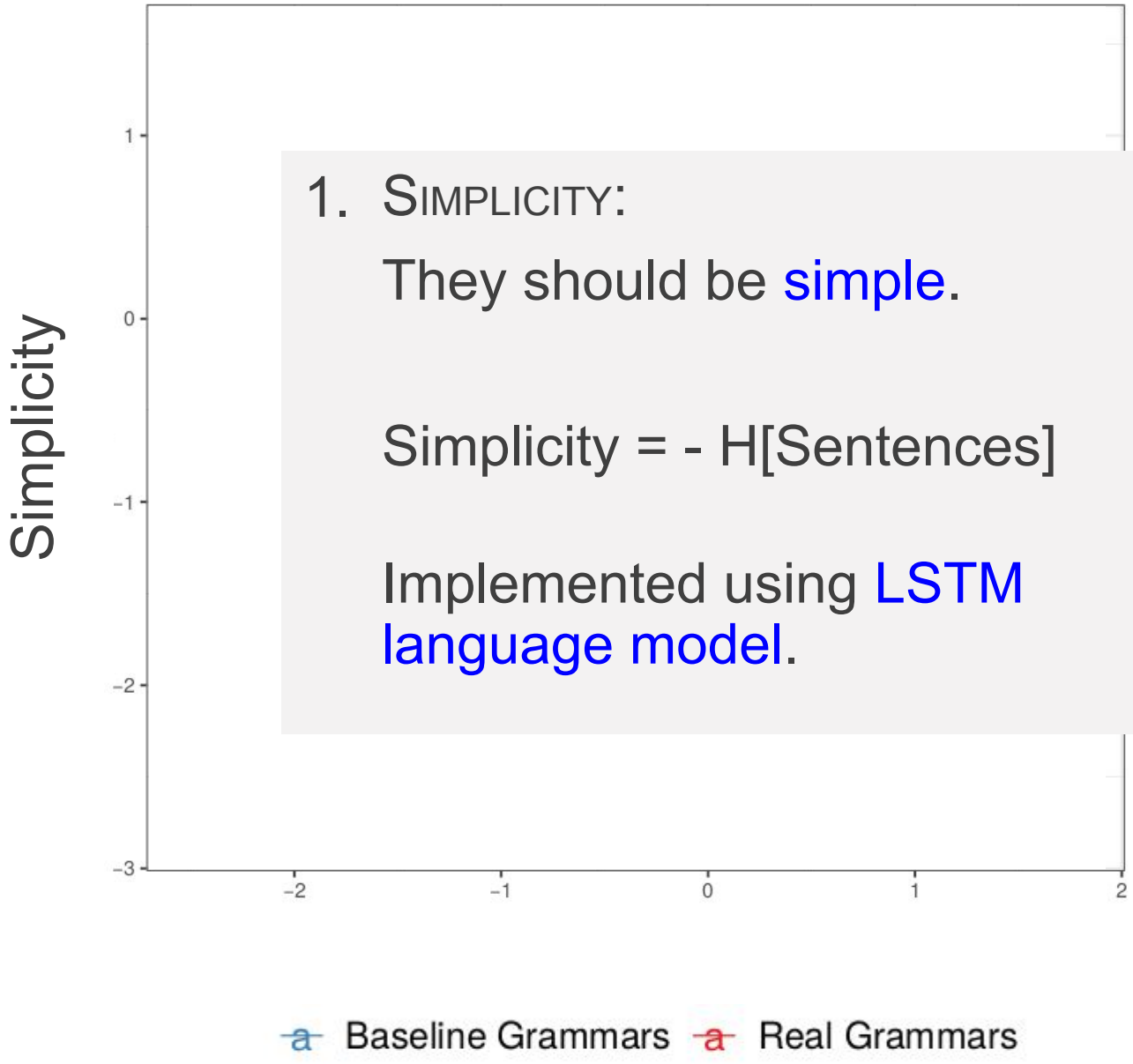


(Dryer 1993, 2013)



(Dryer 2013)


1
0
−1
−2
−3
−2
−1
0
1
2
a Baseline Grammars
a Real Grammars



1. SIMPLICITY:

   They should be simple.

   Simplicity = - H[Sentences]

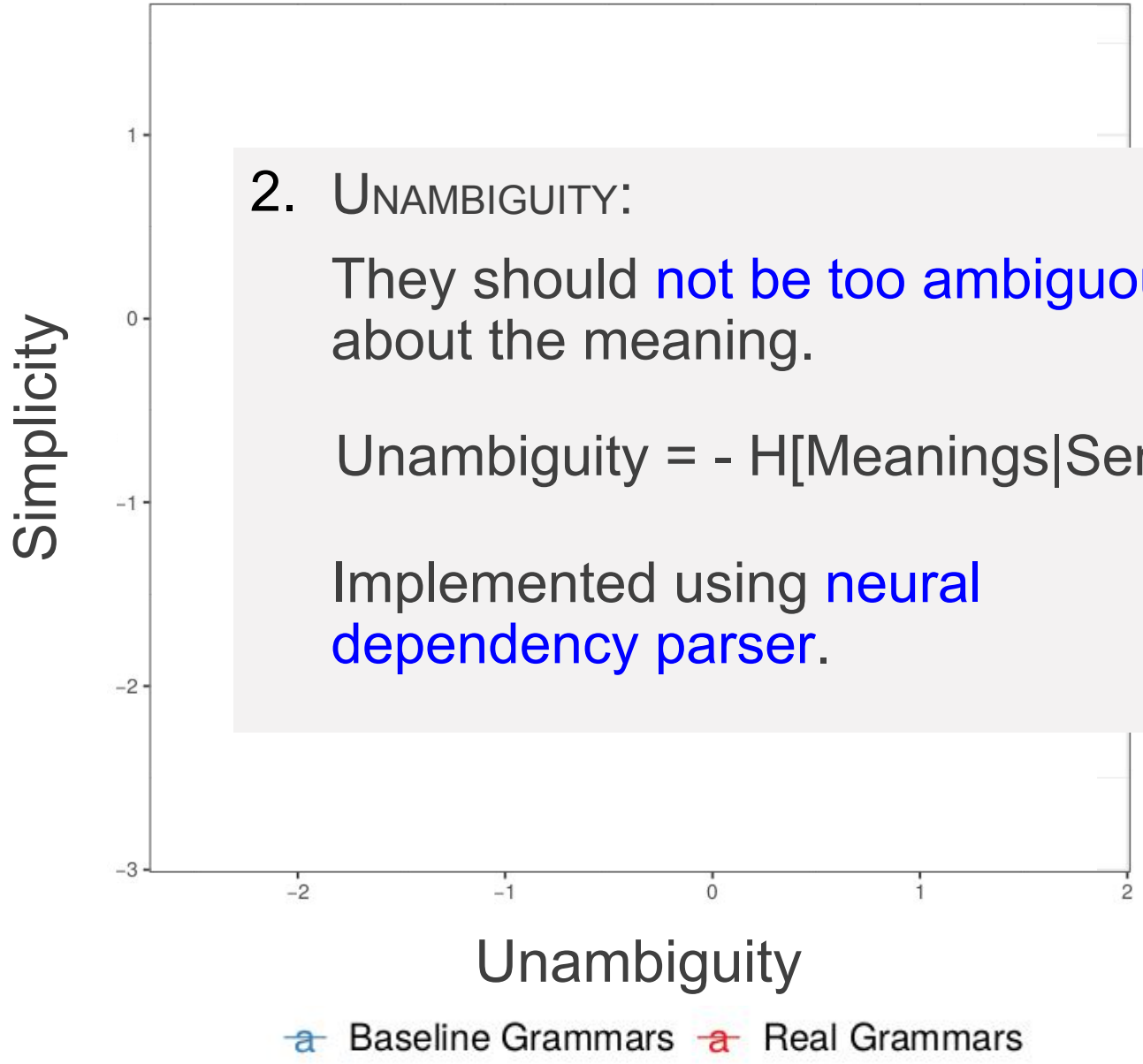
   Implemented using LSTM language model.

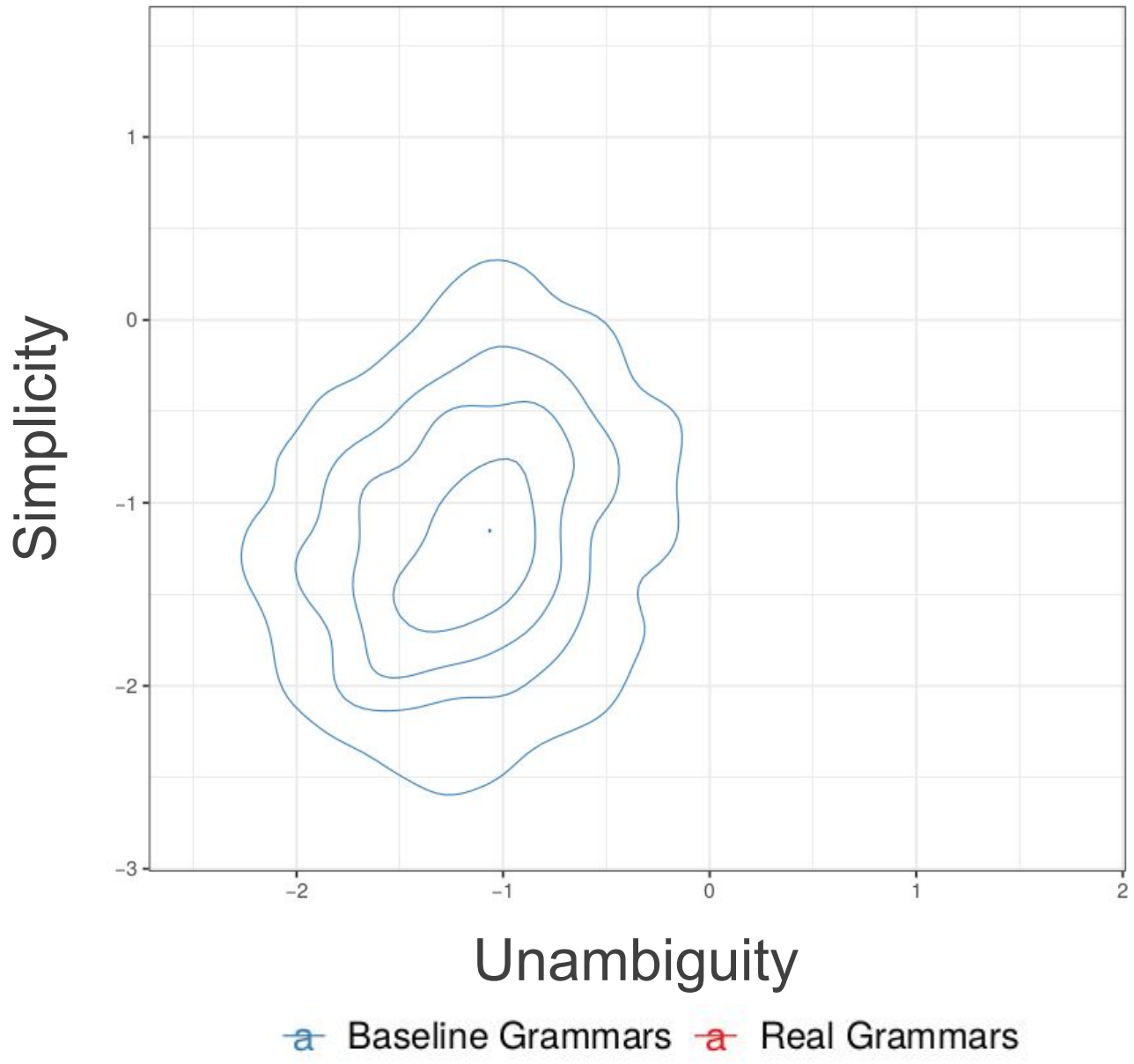
2. UNAMBIGUITY:

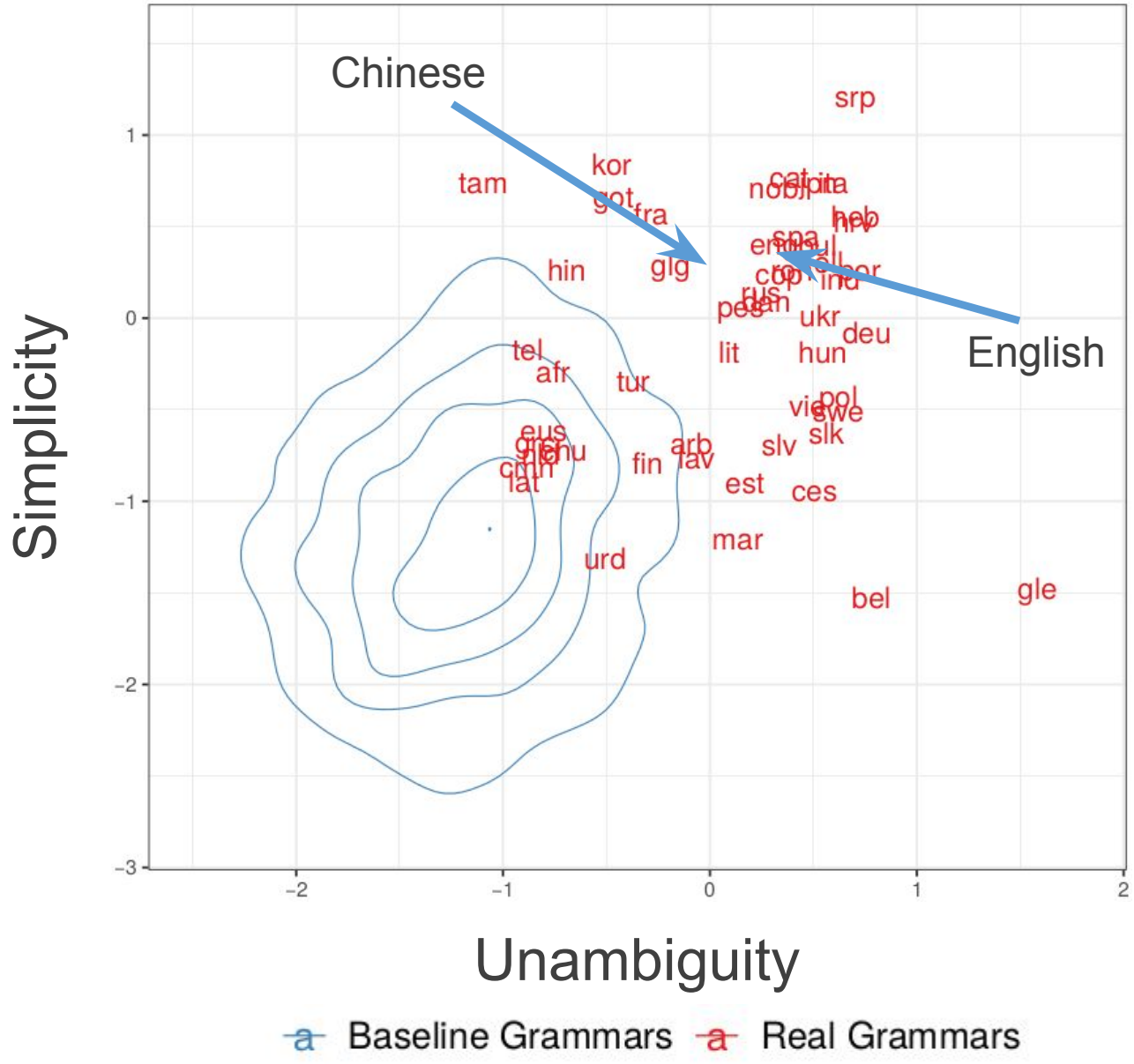
They should not be too ambiguous about the meaning.

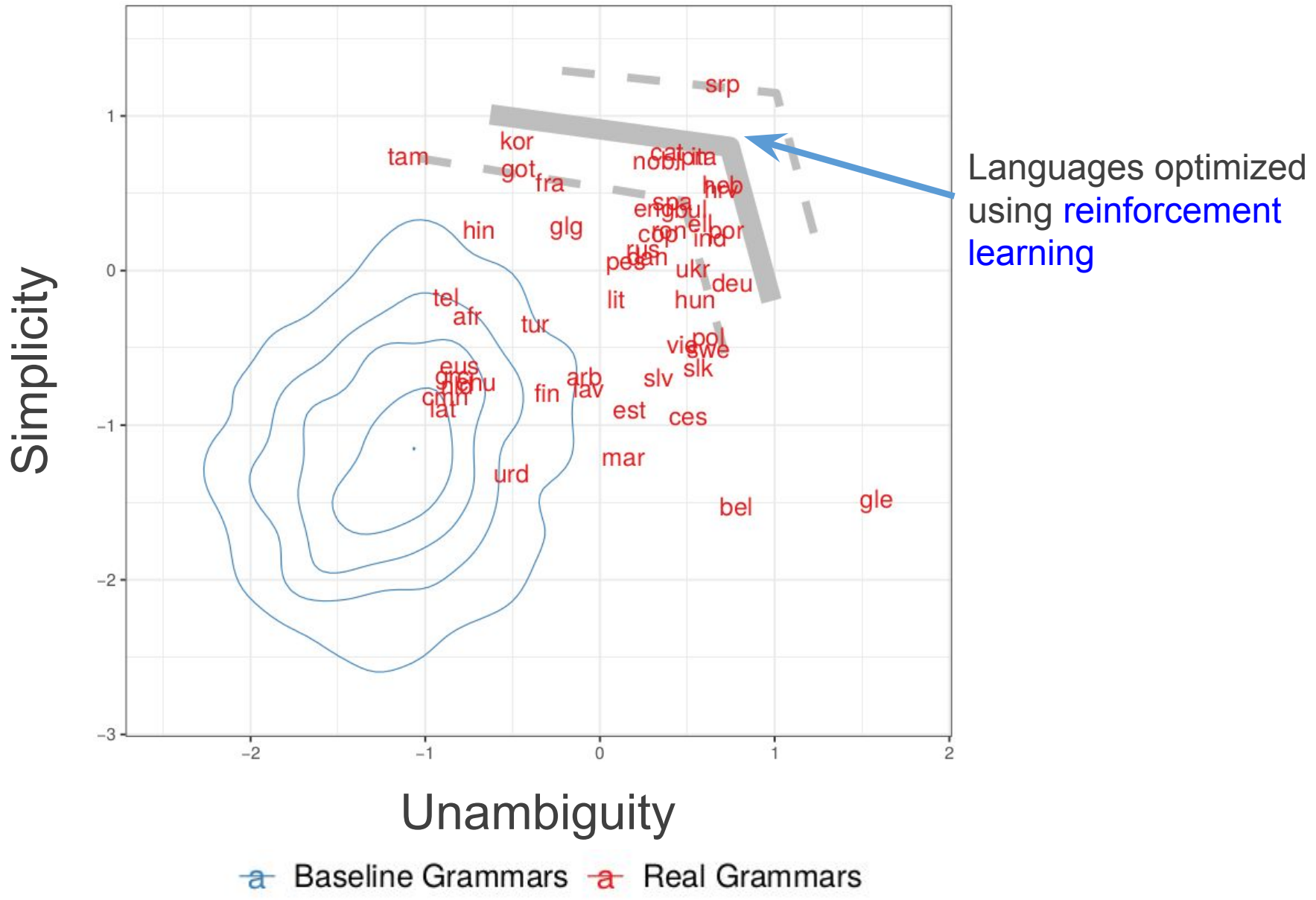
Unambiguity = - H[Meanings|Sentences]

Implemented using neural dependency parser.

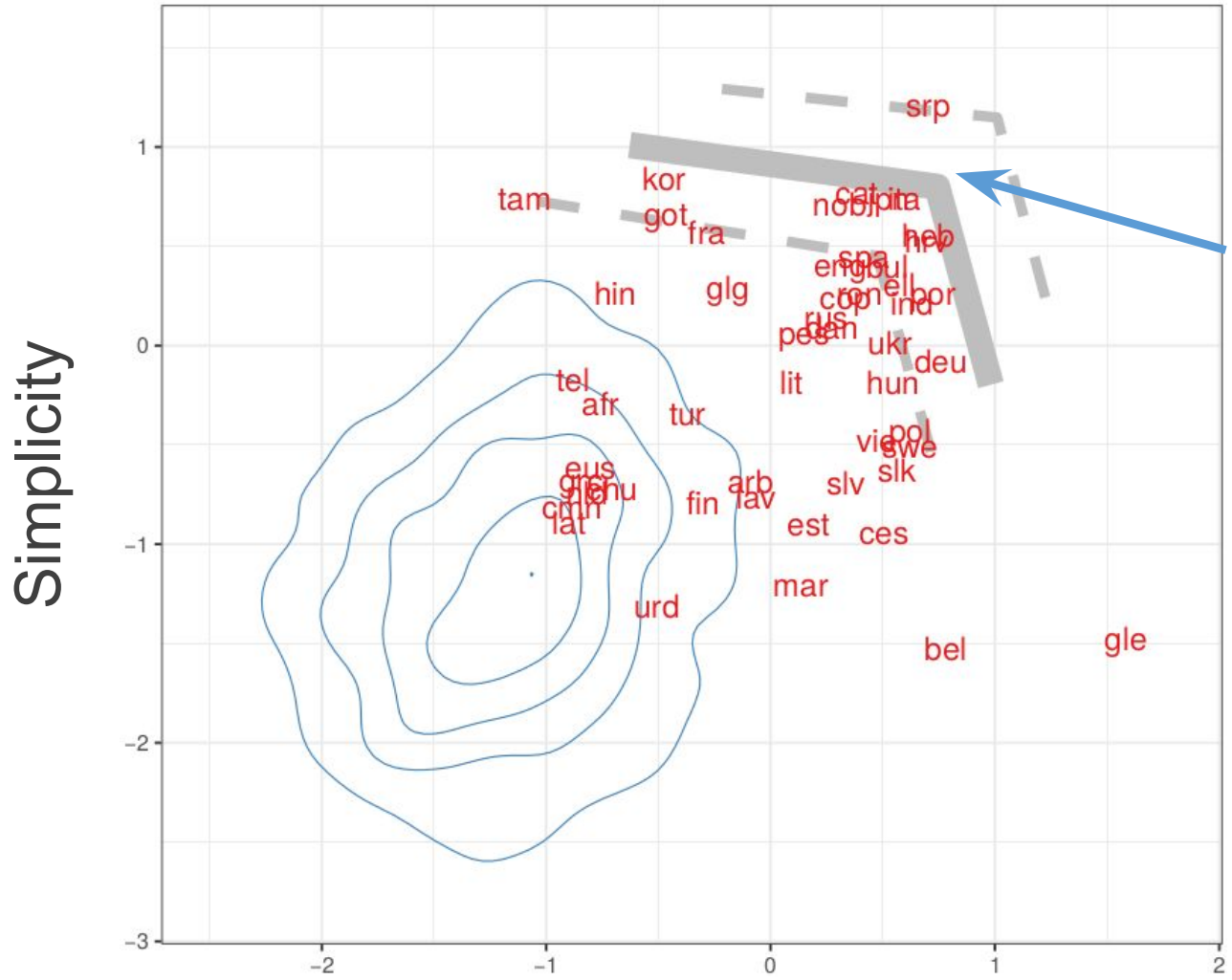



Simplicity
1
0
−1
−2
−3
−2
−1
0
1
2
Unambiguity
Baseline Grammars
Real Grammars


Chinese
English
srp
kor
tam
got
fra
hin
glg
tel
afr
tur
eus
fin
arb
lav
slv
slk
pol
swe
lit
hun
deu
ukr
ces
est
mar
urd
bel
gle
lat
1
0
−1
−2
−3
−2
−1
0
1
2
Simplicity
Unambiguity
Baseline Grammars
Real Grammars

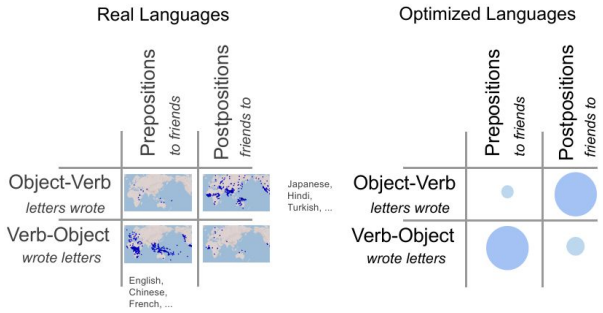


Languages optimized using reinforcement learning

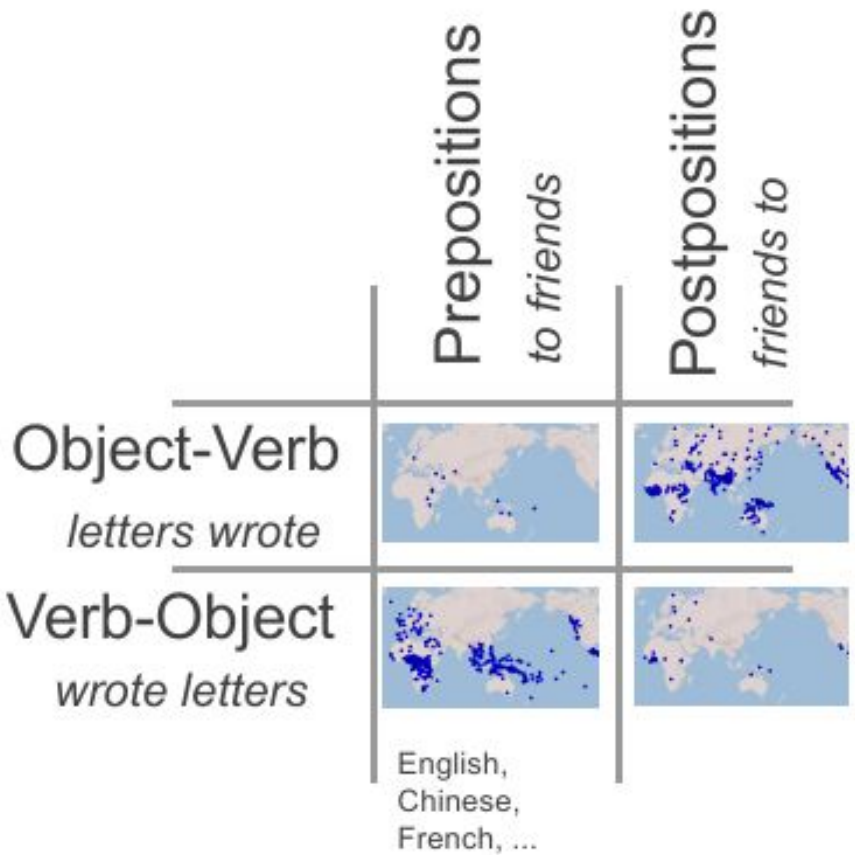


Languages optimized using reinforcement learning

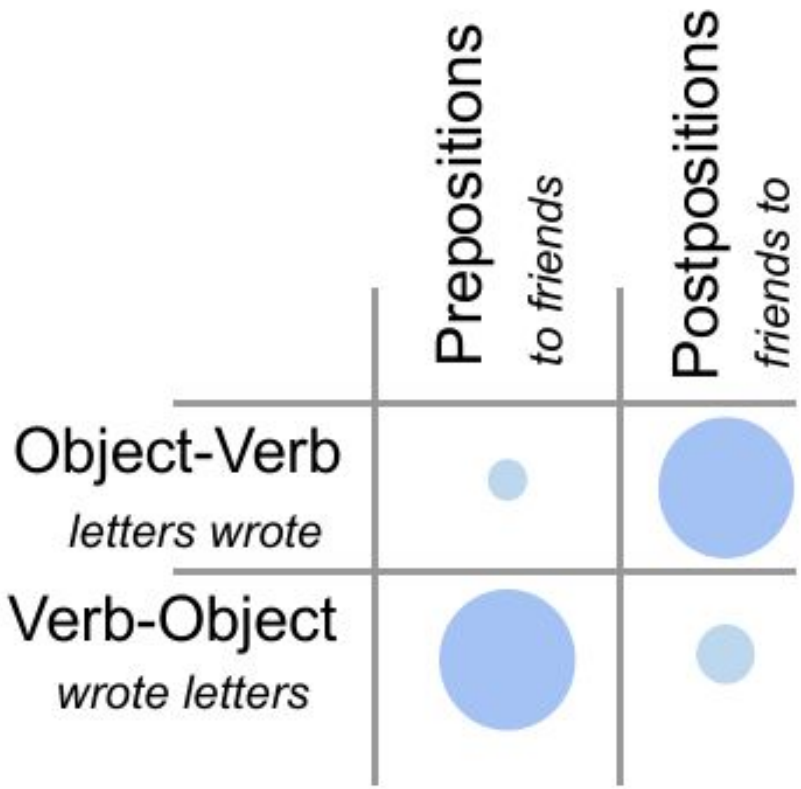
Reproduce many properties of real languages

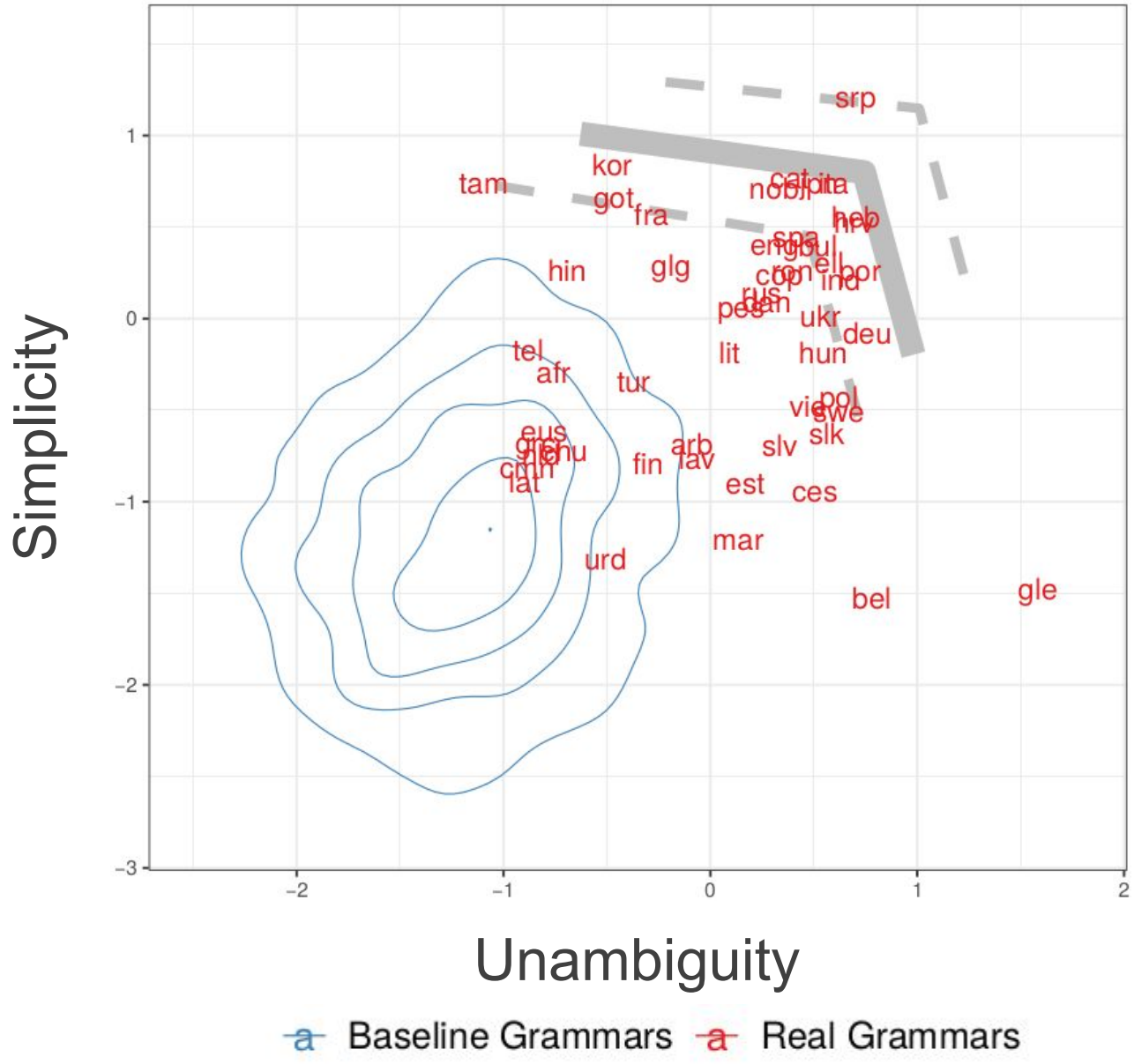


Real Languages



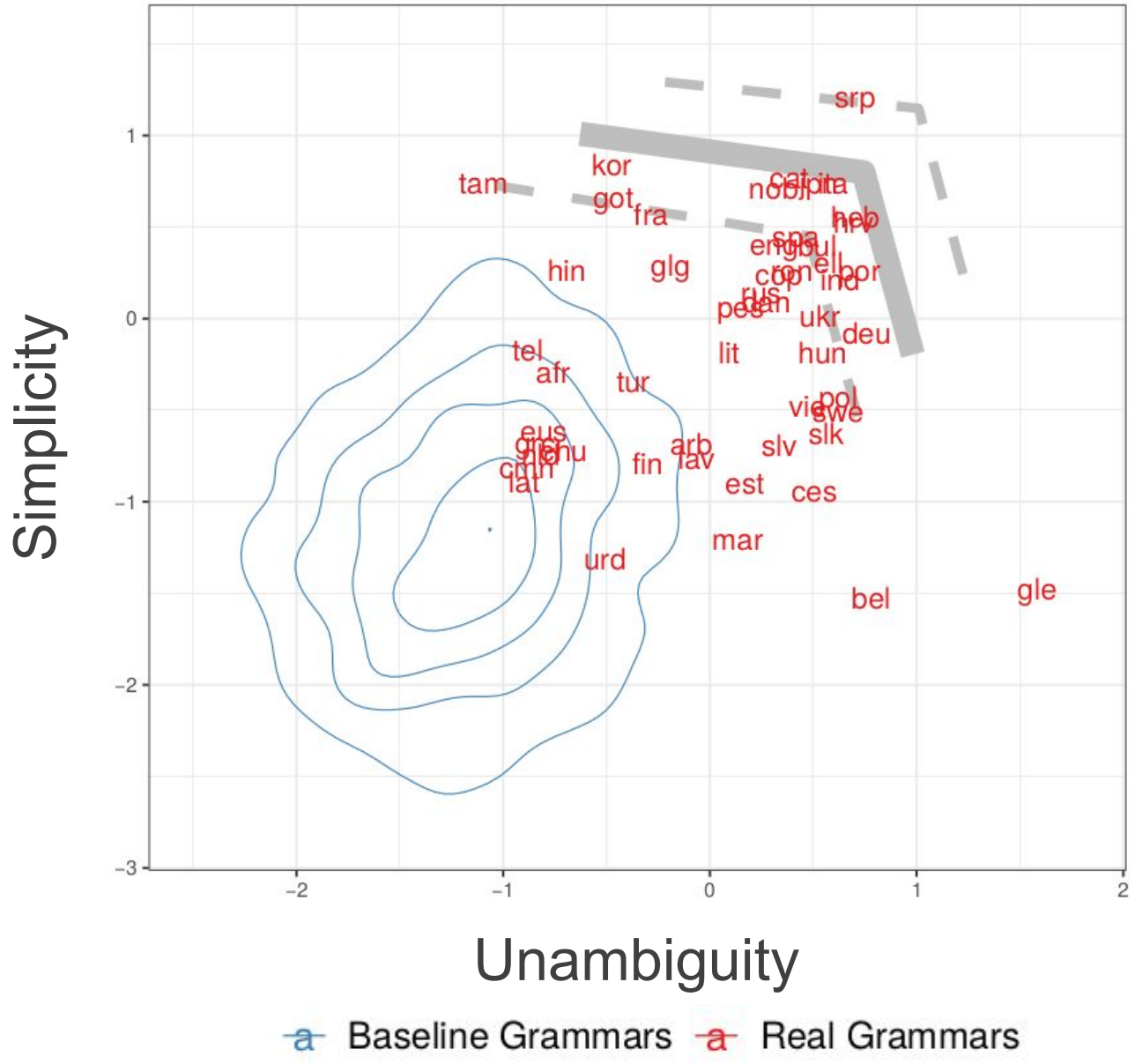
Optimized Languages





Take-Away:

Unsupervised machine learning model predicts properties of human languages



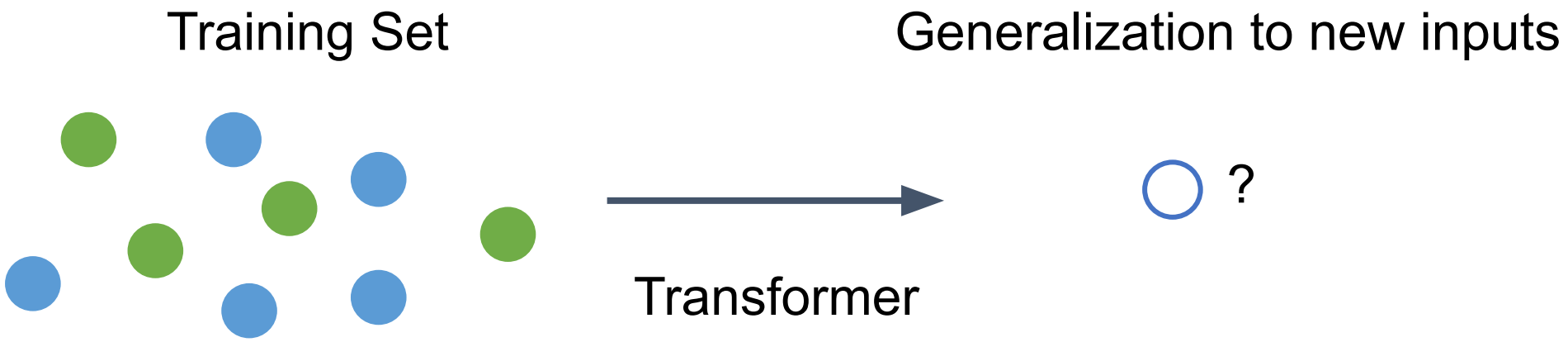
Take-Away:

Unsupervised machine learning model predicts properties of human languages

Could act as prior for NLP models

What's Next?

Characterizing Inductive Bias

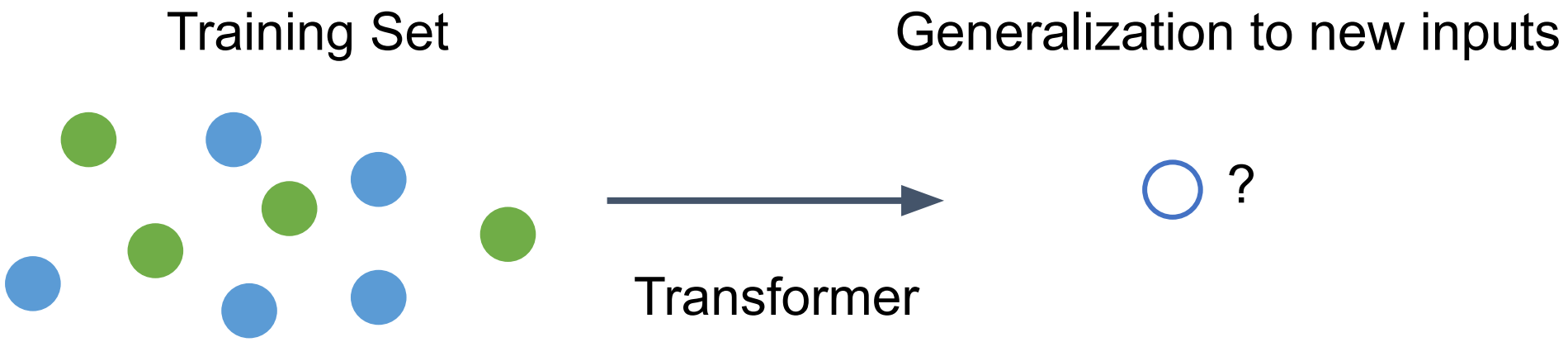


How does the transformer architecture determine generalization?

How does LLM pretraining & finetuning change it?

What's Next?

Characterizing Inductive Bias



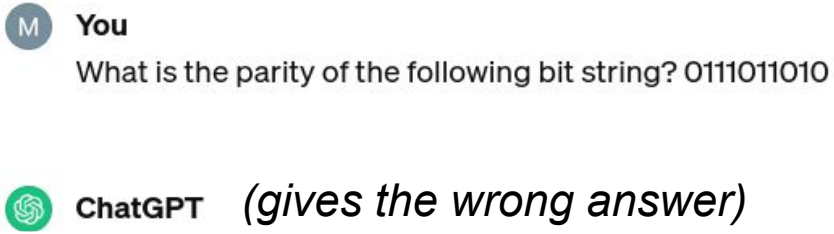
Can we fix wrong LLM answers by targeted training?

What’s Next?

Predicting Difficulty

What's Next?

Predicting Difficulty



What's Next?

Predicting Difficulty

You
What is the parity of the following bit string? 0111011010

ChatGPT *(gives the wrong answer)*

You
Can you do the calculation in more detail step-by-step, counting one-by-one the number of ones?

What's Next?

Predicting Difficulty

You
What is the parity of the following bit string? 0111011010

ChatGPT *(gives the wrong answer)*

You
Can you do the calculation in more detail step-by-step, counting one-by-one the number of ones?

ChatGPT
Certainly! Let's count the number of ones in the bit string 0111011010 step-by-step:

What's Next?

Predicting Difficulty

You
What is the parity of the following bit string? 0111011010

ChatGPT *(gives the wrong answer)*

You
Can you do the calculation in more detail step-by-step, counting one-by-one the number of ones?

ChatGPT
Certainly! Let's count the number of ones in the bit string 0111011010 step-by-step:

less Copy code

```
Step 1: 0 (count: 0)
```

Each step is computationally easier than PARITY.

What's Next?

Predicting Difficulty

You

What is the parity of the following bit string? 0111011010

ChatGPT *(gives the wrong answer)*

You

Can you do the calculation in more detail step-by-step, counting one-by-one the number of ones?

ChatGPT

Certainly! Let's count the number of ones in the bit string 0111011010 step-by-step:

less — Copy code

```
Step 1: 0 (count: 0)
Step 2: 1 (count: 1)
```

Each step is computationally easier than PARITY.

What's Next?

Predicting Difficulty

You

What is the parity of the following bit string? 0111011010

ChatGPT *(gives the wrong answer)*

You

Can you do the calculation in more detail step-by-step, counting one-by-one the number of ones?

ChatGPT

Certainly! Let's count the number of ones in the bit string 0111011010 step-by-step:

less — Copy code

```
Step 1: 0 (count: 0)
Step 2: 1 (count: 1)
Step 3: 1 (count: 2)
```

Each step is computationally easier than PARITY.

What's Next?

Predicting Difficulty

You

What is the parity of the following bit string? 0111011010

ChatGPT *(gives the wrong answer)*

You

Can you do the calculation in more detail step-by-step, counting one-by-one the number of ones?

ChatGPT

Certainly! Let's count the number of ones in the bit string 0111011010 step-by-step:

less — Copy code

```
Step 1: 0 (count: 0)
Step 2: 1 (count: 1)
Step 3: 1 (count: 2)
Step 4: 1 (count: 3)
Step 5: 0 (count: 3)
Step 6: 1 (count: 4)
Step 7: 1 (count: 5)
Step 8: 0 (count: 5)
Step 9: 1 (count: 6)
Step 10: 0 (count: 6)
```

Each step is computationally easier than PARITY.

What's Next?

Predicting Difficulty

ChatGPT

Certainly! Let's count the number of ones in the bit string 0111011010 step-by-step:

less — Copy code

```
Step 1: 0 (count: 0)
Step 2: 1 (count: 1)
Step 3: 1 (count: 2)
Step 4: 1 (count: 3)
Step 5: 0 (count: 3)
Step 6: 1 (count: 4)
Step 7: 1 (count: 5)
Step 8: 0 (count: 5)
Step 9: 1 (count: 6)
Step 10: 0 (count: 6)
```

So, counting one by one, we find that there are 6 ones in the bit string 0111011010. The count is even. Therefore, the parity of the bit string is even.

now, transformer gives the right answer!

What's Next?

Predicting Difficulty

- When does this work?

What's Next?

Predicting Difficulty

When does this work?

Given a problem family, how can we predict how many steps are needed?

What's Next?

Predicting Difficulty

When does this work?

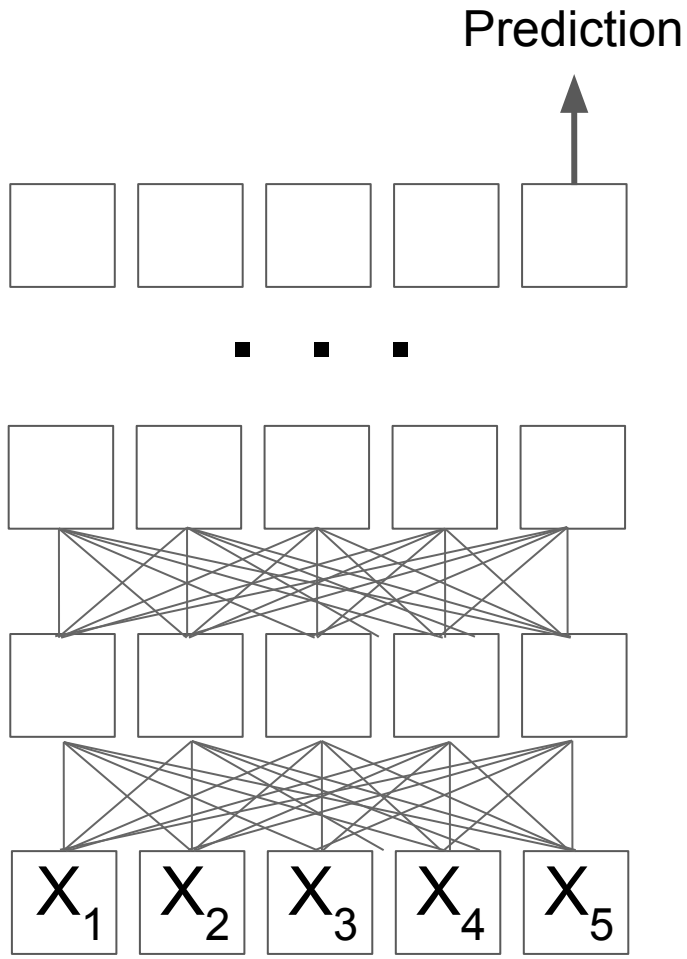
Given a problem family, how can we predict how many steps are needed?

How can we enable to LLM to convincingly certify correctness of each step?

What's Next?

Mechanistic Interpretation

We need techniques for working out how LLMs work internally

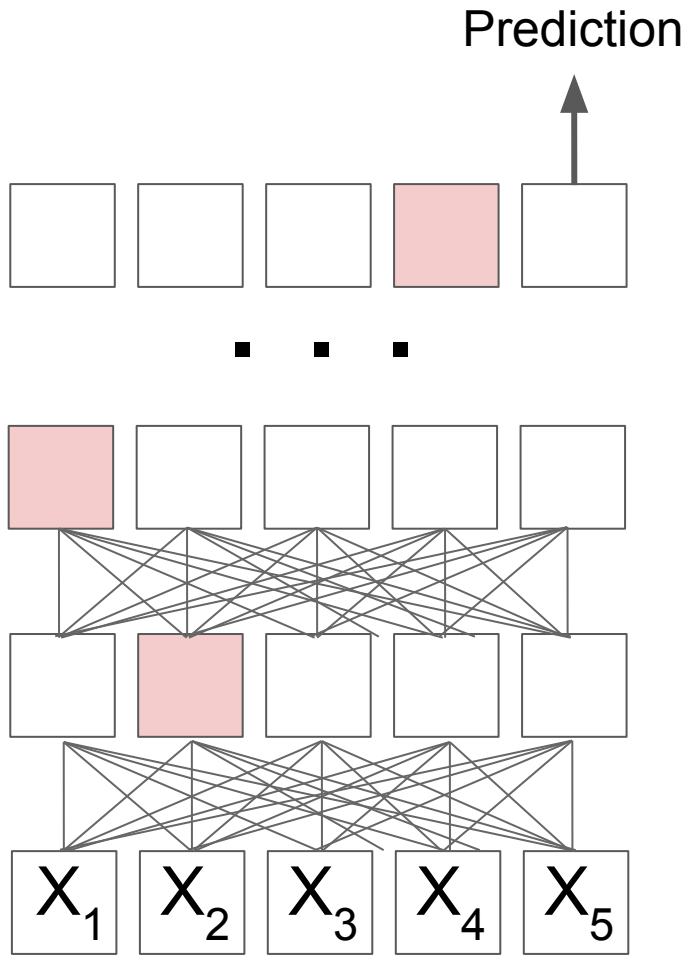


What's Next?

Mechanistic Interpretation

We need techniques for working out how LLMs work internally

Which attention heads attempt to solve parity?

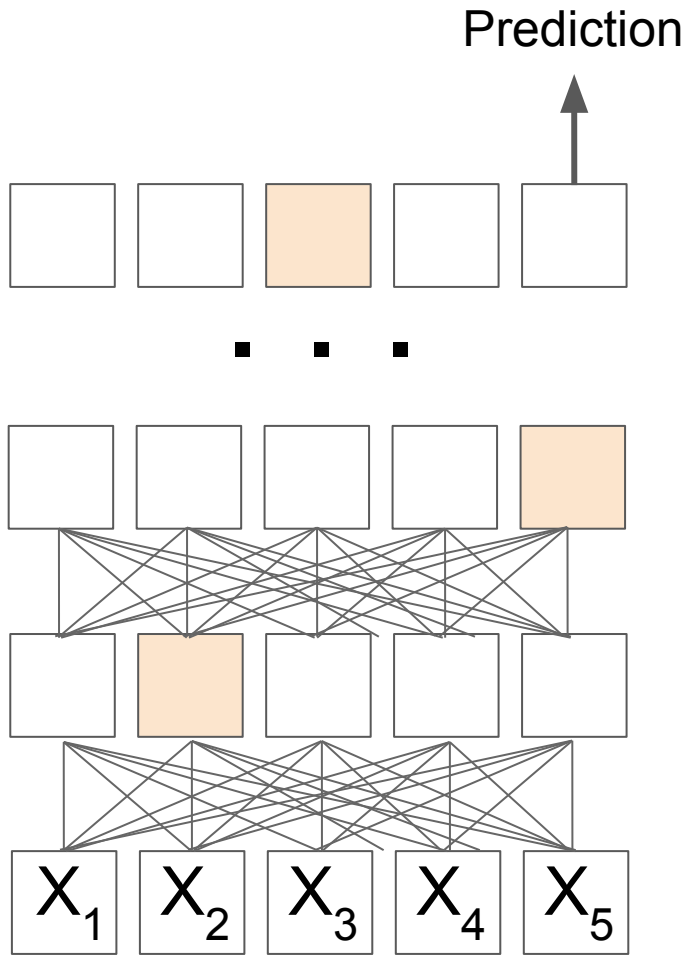


What's Next?

Mechanistic Interpretation

We need techniques for working out how LLMs work internally

Which attention heads attempt to perform parsing?

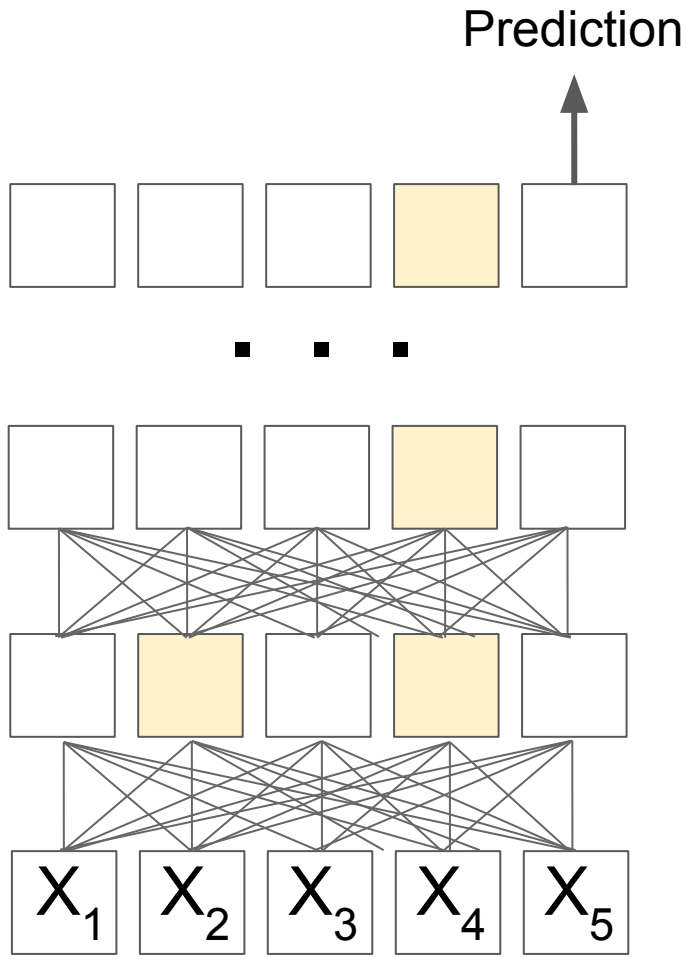


What's Next?

Mechanistic Interpretation

We need techniques for working out how LLMs work internally

Which attention heads extract knowledge from context?

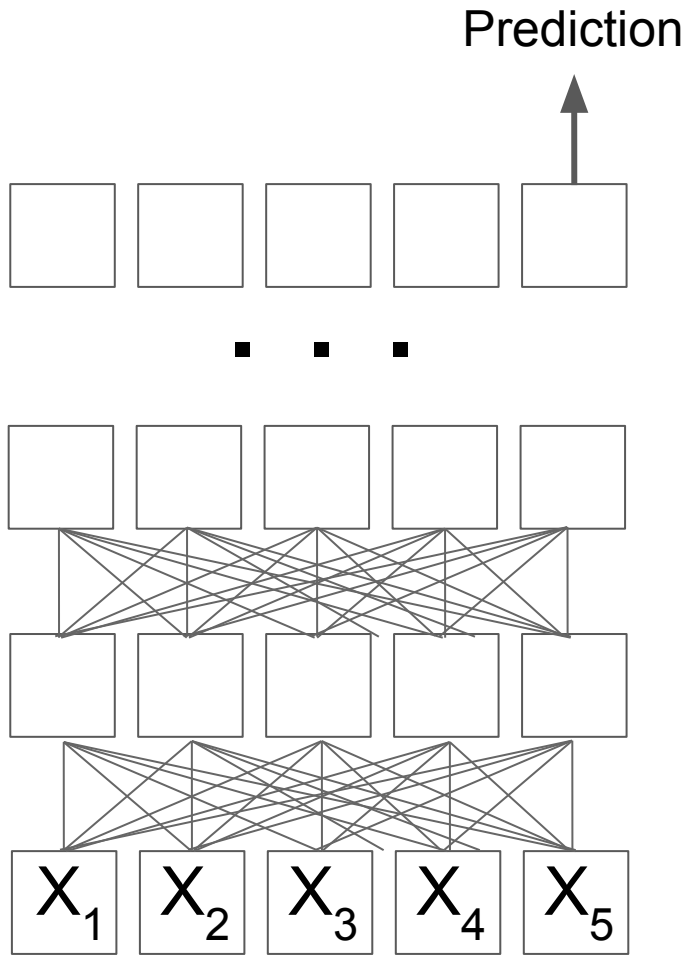


What's Next?

Mechanistic Interpretation

We need techniques for working out how LLMs work internally

Will enable better understanding of failures.



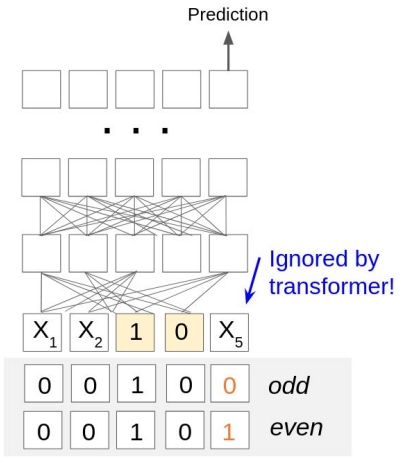
Conclusion

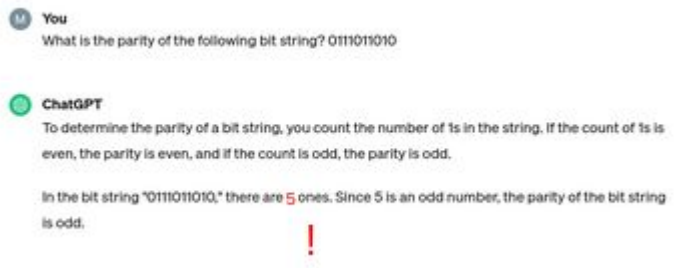
Capabilities and limitations of neural ML architectures

Reverse-engineer human language comprehension using LMs

Conclusion

Capabilities and limitations of neural ML architectures

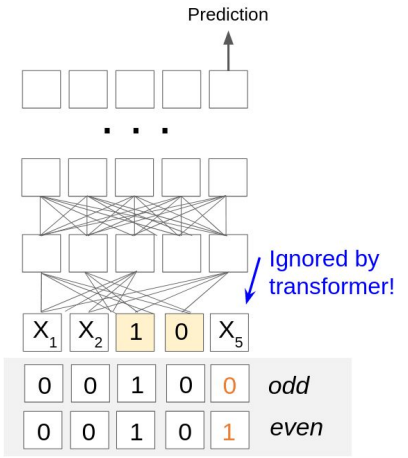


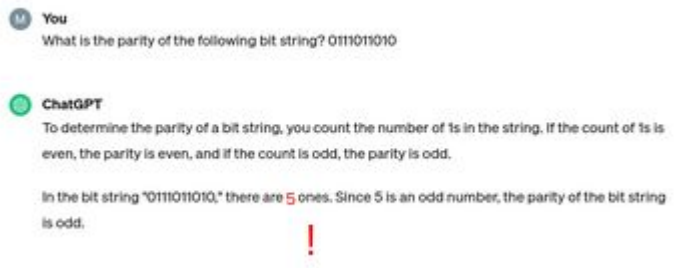


Reverse-engineer human language comprehension using LMs

Conclusion

Capabilities and limitations of neural ML architectures

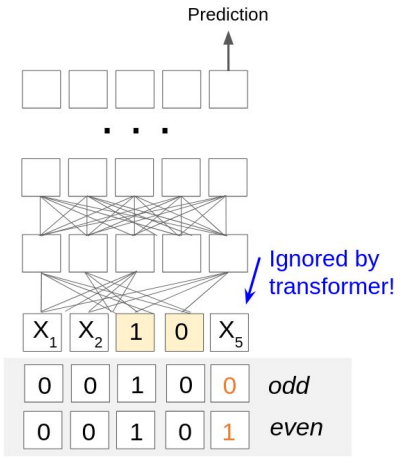


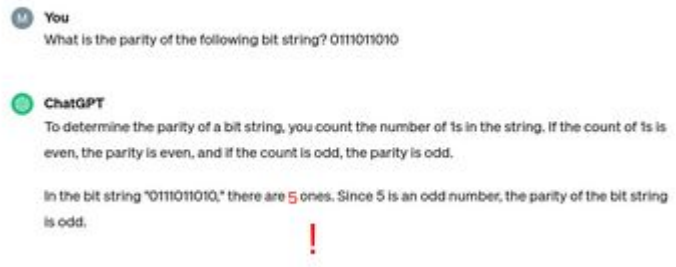


Reverse-engineer human language comprehension using LMs

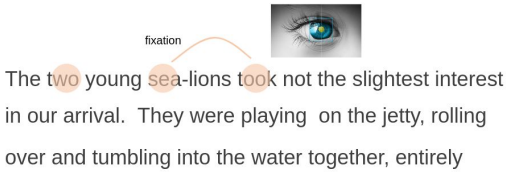
Conclusion

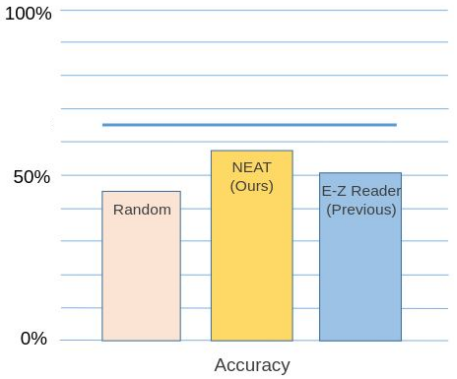
Capabilities and limitations of neural ML architectures





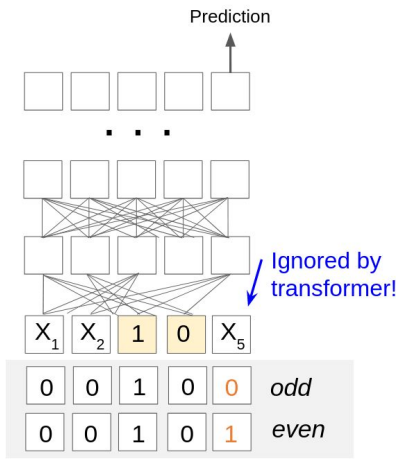
Reverse-engineer human language comprehension using LMs

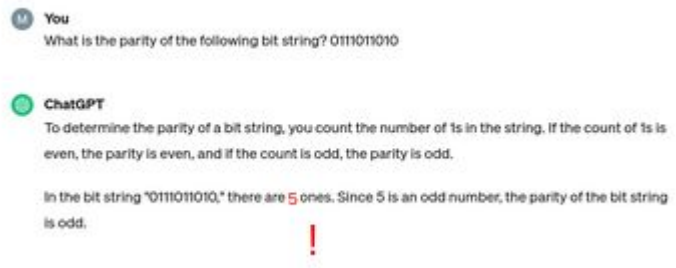




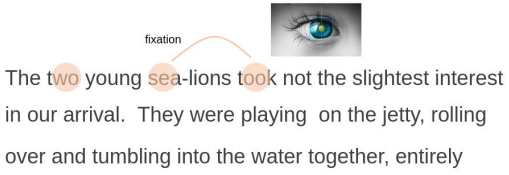
Conclusion

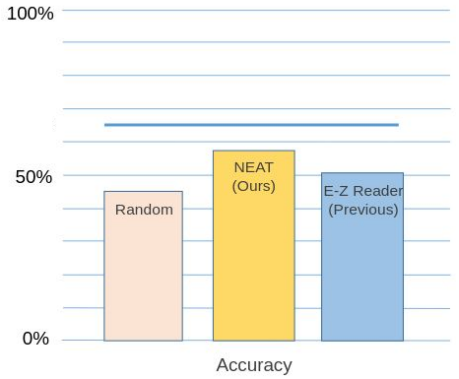
Capabilities and limitations of neural ML architectures





Reverse-engineer human language comprehension using LMs





Thank you!