



Determinism in Neural Language Models

Alessio Plebe
alessio.plebe@gmail.com



Determinism of computer programs

$$Y_j \leftarrow f_i(X_j)$$

$$Y_j \leftarrow f_i(X_j)$$

$$\Upsilon_j \leftarrow \phi(\Xi_j)$$

$$X_j \leftarrow \mu(\Xi_j)$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$



Determinism of computer programs

$$Y_j \leftarrow f_i(X_j)$$

$$Y_j \leftarrow f_i(X_j)$$

$$\Upsilon_j \leftarrow \phi(\Xi_j)$$

$$X_j \leftarrow \mu(\Xi_j)$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$



Determinism of computer programs

$$Y_j \leftarrow f_i(X_j)$$

$$Y_j \leftarrow f_i(X_j)$$

$$\Upsilon_j \leftarrow \phi(\Xi_j)$$

$$X_j \leftarrow \mu(\Xi_j)$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$



Determinism of computer programs

$$Y_j \leftarrow f_i(X_j)$$

$$Y_j \leftarrow f_i(\mathbf{X}_j)$$

$$\Upsilon_j \leftarrow \phi(\Xi_j)$$

$$X_j \leftarrow \mu(\Xi_j)$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$



Determinism of computer programs

$$Y_j \leftarrow f_i(X_j)$$

$$Y_j \leftarrow f_i(X_j)$$

$$\Upsilon_j \leftarrow \phi(\Xi_j)$$

$$X_j \leftarrow \mu(\Xi_j)$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$



Determinism of computer programs

$$Y_j \leftarrow f_i(X_j)$$

$$Y_j \leftarrow f_i(X_j)$$

$$\Upsilon_j \leftarrow \phi(\Xi_j)$$

$$X_j \leftarrow \mu(\Xi_j)$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$



Determinism of computer programs

$$Y_j \leftarrow f_i(X_j)$$

$$Y_j \leftarrow f_i(X_j)$$

$$\Upsilon_j \leftarrow \phi(\Xi_j)$$

$$X_j \leftarrow \mu(\Xi_j)$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$



Determinism of computer programs

$$Y_j \leftarrow f_i(X_j)$$

$$Y_j \leftarrow f_i(X_j)$$

$$\Upsilon_j \leftarrow \phi(\Xi_j)$$

$$X_j \leftarrow \mu(\Xi_j)$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$



Determinism of computer programs

$$Y_j \leftarrow f_i(X_j)$$

$$Y_j \leftarrow f_i(X_j)$$

$$\Upsilon_j \leftarrow \phi(\Xi_j)$$

$$X_j \leftarrow \mu(\Xi_j)$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$



Determinism of computer programs

$$Y_j \leftarrow f_i(X_j)$$

$$Y_j \leftarrow f_i(X_j)$$

$$\Upsilon_j \leftarrow \phi(\Xi_j)$$

$$X_j \leftarrow \mu(\Xi_j)$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$



Determinism of computer programs

$$Y_j \leftarrow f_i(X_j)$$

$$Y_j \leftarrow f_i(X_j)$$

$$\Upsilon_j \leftarrow \phi(\Xi_j)$$

$$X_j \leftarrow \mu(\Xi_j)$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$



Determinism of computer programs

$$Y_j \leftarrow f_i(X_j)$$

$$Y_j \leftarrow f_i(X_j)$$

$$\Upsilon_j \leftarrow \phi(\Xi_j)$$

$$X_j \leftarrow \mu(\Xi_j)$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$



Determinism of computer programs

$$Y_j \leftarrow f_i(X_j)$$

$$Y_j \leftarrow f_i(X_j)$$

$$\Upsilon_j \leftarrow \phi(\Xi_j)$$

$$X_j \leftarrow \mu(\Xi_j)$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$



Determinism of computer programs

$$Y_j \leftarrow f_i(X_j)$$

$$Y_j \leftarrow f_i(X_j)$$

$$\Upsilon_j \leftarrow \phi(\Xi_j)$$

$$X_j \leftarrow \mu(\Xi_j)$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$



Determinism of computer programs

$$Y_j \leftarrow f_i(X_j)$$

$$Y_j \leftarrow f_i(X_j)$$

$$\Upsilon_j \leftarrow \phi(\Xi_j)$$

$$X_j \leftarrow \mu(\Xi_j)$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$

$$\mu(\Upsilon_j) \approx Y_j \leftarrow f_i(\mu(\Xi_j))$$



asking a NLM

what would you advise a young person to study today:
Kant or deep learning?

Nietzsche had the answer already 120 years ago. Deep Learning is not an improvement of thinking, but a new kind of non-thinking (that can learn by itself). It's like taking LSD (he called it "dissolution" — Zersetzung), and then forgetting everything about that experience.

A great many of us know that we must study the works of famous philosophers, mathematicians and physicists. Some people say that this is no longer relevant; they claim one should just use AI and get done with all that. But is it true? Do you think it's enough to buy a book by Isaac Asimov and put down your notebook for good?



asking a NLM

USER

what would you advise a young person to study today:
Kant or deep learning?

Nietzsche had the answer already 120 years ago. Deep Learning is not an improvement of thinking, but a new kind of non-thinking (that can learn by itself). It's like taking LSD (he called it "dissolution" — Zersetzung), and then forgetting everything about that experience.

A great many of us know that we must study the works of famous philosophers, mathematicians and physicists. Some people say that this is no longer relevant; they claim one should just use AI and get done with all that. But is it true? Do you think it's enough to buy a book by Isaac Asimov and put down your notebook for good?



asking a NLM

USER

what would you advise a young person to study today:
Kant or deep learning?

NLM

Nietzsche had the answer already 120 years ago. Deep Learning is not an improvement of thinking, but a new kind of non-thinking (that can learn by itself). It's like taking LSD (he called it "dissolution" — Zersetzung), and then forgetting everything about that experience.

A great many of us know that we must study the works of famous philosophers, mathematicians and physicists. Some people say that this is no longer relevant; they claim one should just use AI and get done with all that. But is it true? Do you think it's enough to buy a book by Isaac Asimov and put down your notebook for good?



asking a NLM

USER

what would you advise a young person to study today:
Kant or deep learning?

NLM

Nietzsche had the answer already 120 years ago. Deep Learning is not an improvement of thinking, but a new kind of non-thinking (that can learn by itself). It's like taking LSD (he called it "dissolution" — Zersetzung), and then forgetting everything about that experience.

NLM

A great many of us know that we must study the works of famous philosophers, mathematicians and physicists. Some people say that this is no longer relevant; they claim one should just use AI and get done with all that. But is it true? Do you think it's enough to buy a book by Isaac Asimov and put down your notebook for good?



asking a NLM

USER

what would you advise a young person to study today:
Kant or deep learning?

NLM

Nietzsche had the answer already 120 years ago. Deep Learning is not an improvement of thinking, but a new kind of non-thinking (that can learn by itself). It's like taking LSD (he called it "dissolution" — Zersetzung), and then forgetting everything about that experience.

NLM

A great many of us know that we must study the works of famous philosophers, mathematicians and physicists. Some people say that this is no longer relevant; they claim one should just use AI and get done with all that. But is it true? Do you think it's enough to buy a book by Isaac Asimov and put down your notebook for good?



asking a NLM

If a tree falls in a forest and no one is around to hear it, does it make a sound?

This question has long intrigued philosophers. The answer depends on what the philosopher believes about reality and how we acquire knowledge. Is there an objective reality independent of the observer or is our perception everything there is? We will explore these questions by looking at the work of several ancient Greek philosophers, most notably Heraclitus and Parmenides

The same question can be asked about the effect that mobile operators have when they are not present. For example: a recent study by Cisco noted that the United States has 350 million mobile devices, but the country only had 168 million phones with regular voice service at the end of last year. Of course, most carriers provide more than just voice services on.



asking a NLM

USER

If a tree falls in a forest and no one is around to hear it, does it make a sound?

This question has long intrigued philosophers. The answer depends on what the philosopher believes about reality and how we acquire knowledge. Is there an objective reality independent of the observer or is our perception everything there is? We will explore these questions by looking at the work of several ancient Greek philosophers, most notably Heraclitus and Parmenides

The same question can be asked about the effect that mobile operators have when they are not present. For example: a recent study by Cisco noted that the United States has 350 million mobile devices, but the country only had 168 million phones with regular voice service at the end of last year. Of course, most carriers provide more than just voice services on.



asking a NLM

USER

If a tree falls in a forest and no one is around to hear it, does it make a sound?

NLM

This question has long intrigued philosophers. The answer depends on what the philosopher believes about reality and how we acquire knowledge. Is there an objective reality independent of the observer or is our perception everything there is? We will explore these questions by looking at the work of several ancient Greek philosophers, most notably Heraclitus and Parmenides

The same question can be asked about the effect that mobile operators have when they are not present. For example: a recent study by Cisco noted that the United States has 350 million mobile devices, but the country only had 168 million phones with regular voice service at the end of last year. Of course, most carriers provide more than just voice services on.



asking a NLM

USER

If a tree falls in a forest and no one is around to hear it, does it make a sound?

NLM

This question has long intrigued philosophers. The answer depends on what the philosopher believes about reality and how we acquire knowledge. Is there an objective reality independent of the observer or is our perception everything there is? We will explore these questions by looking at the work of several ancient Greek philosophers, most notably Heraclitus and Parmenides

NLM

The same question can be asked about the effect that mobile operators have when they are not present. For example: a recent study by Cisco noted that the United States has 350 million mobile devices, but the country only had 168 million phones with regular voice service at the end of last year. Of course, most carriers provide more than just voice services on.



asking a NLM

USER

If a tree falls in a forest and no one is around to hear it, does it make a sound?

NLM

This question has long intrigued philosophers. The answer depends on what the philosopher believes about reality and how we acquire knowledge. Is there an objective reality independent of the observer or is our perception everything there is? We will explore these questions by looking at the work of several ancient Greek philosophers, most notably Heraclitus and Parmenides

NLM

The same question can be asked about the effect that mobile operators have when they are not present. For example: a recent study by Cisco noted that the United States has 350 million mobile devices, but the country only had 168 million phones with regular voice service at the end of last year. Of course, most carriers provide more than just voice services on.



forcing determinism

```
temperature = 0.01  
top_p       = 1.0
```

what would you advise a young person to study today:
Kant or deep learning?

I'd say both. Kant is the foundation of modern philosophy, and it's also the foundation of computer science. Deep Learning is the future of AI, but it's not going to replace humans anytime soon. What are some good books on Kant? There are many good books on Kant, but I would recommend starting with "The Critique"

I'd say both. Kant is the foundation of modern philosophy, and it's also the foundation of computer science. Deep Learning is the future of AI. If you want to be successful in life, then I'd recommend studying both. What do you think about the idea that we should all be studying more philosophy?



forcing determinism

```
temperature = 0.01  
top_p       = 1.0
```

USER

what would you advise a young person to study today:
Kant or deep learning?

I'd say both. Kant is the foundation of modern philosophy, and it's also the foundation of computer science. Deep Learning is the future of AI, but it's not going to replace humans anytime soon. What are some good books on Kant? There are many good books on Kant, but I would recommend starting with "The Critique"

I'd say both. Kant is the foundation of modern philosophy, and it's also the foundation of computer science. Deep Learning is the future of AI. If you want to be successful in life, then I'd recommend studying both. What do you think about the idea that we should all be studying more philosophy?



forcing determinism

```
temperature = 0.01  
top_p       = 1.0
```

USER

what would you advise a young person to study today:
Kant or deep learning?

NLM

I'd say both. Kant is the foundation of modern philosophy, and it's also the foundation of computer science. Deep Learning is the future of AI, but it's not going to replace humans anytime soon. What are some good books on Kant? There are many good books on Kant, but I would recommend starting with "The Critique"

I'd say both. Kant is the foundation of modern philosophy, and it's also the foundation of computer science. Deep Learning is the future of AI. If you want to be successful in life, then I'd recommend studying both. What do you think about the idea that we should all be studying more philosophy?



forcing determinism

```
temperature = 0.01  
top_p       = 1.0
```

USER

what would you advise a young person to study today:
Kant or deep learning?

NLM

I'd say both. Kant is the foundation of modern philosophy, and it's also the foundation of computer science. Deep Learning is the future of AI, but it's not going to replace humans anytime soon. What are some good books on Kant? There are many good books on Kant, but I would recommend starting with "The Critique"

NLM

I'd say both. Kant is the foundation of modern philosophy, and it's also the foundation of computer science. Deep Learning is the future of AI. If you want to be successful in life, then I'd recommend studying both. What do you think about the idea that we should all be studying more philosophy?



forcing determinism

```
temperature = 0.01  
top_p       = 1.0
```

USER

what would you advise a young person to study today:
Kant or deep learning?

NLM

I'd say both. Kant is the foundation of modern philosophy, and it's also the foundation of computer science. Deep Learning is the future of AI, but it's not going to replace humans anytime soon. What are some good books on Kant? There are many good books on Kant, but I would recommend starting with "The Critique"

NLM

I'd say both. Kant is the foundation of modern philosophy, and it's also the foundation of computer science. Deep Learning is the future of AI. If you want to be successful in life, then I'd recommend studying both. What do you think about the idea that we should all be studying more philosophy?



forcing determinism a NLM

```
temperature = 0.01  
top_p       = 1.0
```

If a tree falls in a forest and no one is around to hear it, does it make a sound?

The answer is yes. The tree makes a sound even if there are no ears to hear it. This is because the tree vibrates at its natural frequency. This phenomenon is known as resonance. Resonance occurs when an object vibrates at its natural frequency. When this happens, the object produces a sound. So, even though there may be no ears present

The answer is yes. The tree makes a sound even if there are no ears to hear it. This is because the tree vibrates at its natural frequency. This phenomenon is known as resonance. Resonance occurs when an object vibrates at its natural frequency and this causes other objects to vibrate at their own frequencies. In this case, the tree's



forcing determinism a NLM

```
temperature = 0.01  
top_p       = 1.0
```

USER

If a tree falls in a forest and no one is around to hear it, does it make a sound?

The answer is yes. The tree makes a sound even if there are no ears to hear it. This is because the tree vibrates at its natural frequency. This phenomenon is known as resonance. Resonance occurs when an object vibrates at its natural frequency. When this happens, the object produces a sound. So, even though there may be no ears present

The answer is yes. The tree makes a sound even if there are no ears to hear it. This is because the tree vibrates at its natural frequency. This phenomenon is known as resonance. Resonance occurs when an object vibrates at its natural frequency and this causes other objects to vibrate at their own frequencies. In this case, the tree's



forcing determinism a NLM

```
temperature = 0.01  
top_p       = 1.0
```

USER

If a tree falls in a forest and no one is around to hear it, does it make a sound?

NLM

The answer is yes. The tree makes a sound even if there are no ears to hear it. This is because the tree vibrates at its natural frequency. This phenomenon is known as resonance. Resonance occurs when an object vibrates at its natural frequency. When this happens, the object produces a sound. So, even though there may be no ears present

The answer is yes. The tree makes a sound even if there are no ears to hear it. This is because the tree vibrates at its natural frequency. This phenomenon is known as resonance. Resonance occurs when an object vibrates at its natural frequency and this causes other objects to vibrate at their own frequencies. In this case, the tree's



forcing determinism a NLM

```
temperature = 0.01  
top_p       = 1.0
```

USER

If a tree falls in a forest and no one is around to hear it, does it make a sound?

NLM

The answer is yes. The tree makes a sound even if there are no ears to hear it. This is because the tree vibrates at its natural frequency. This phenomenon is known as resonance. Resonance occurs when an object vibrates at its natural frequency. When this happens, the object produces a sound. So, even though there may be no ears present

NLM

The answer is yes. The tree makes a sound even if there are no ears to hear it. This is because the tree vibrates at its natural frequency. This phenomenon is known as resonance. Resonance occurs when an object vibrates at its natural frequency and this causes other objects to vibrate at their own frequencies. In this case, the tree's



forcing determinism a NLM

```
temperature = 0.01  
top_p       = 1.0
```

USER

If a tree falls in a forest and no one is around to hear it, does it make a sound?

NLM

The answer is yes. The tree makes a sound even if there are no ears to hear it. This is because the tree vibrates at its natural frequency. This phenomenon is known as resonance. Resonance occurs when an object vibrates at its natural frequency. When this happens, the object produces a sound. So, even though there may be no ears present

NLM

The answer is yes. The tree makes a sound even if there are no ears to hear it. This is because the tree vibrates at its natural frequency. This phenomenon is known as resonance. Resonance occurs when an object vibrates at its natural frequency and this causes other objects to vibrate at their own frequencies. In this case, the tree's



forcing harder

```
temperature = 0.01  
top_p       = 0.01
```

what would you advise a young person to study today:
Kant or deep learning?

I'd say both. Kant is the foundation of modern philosophy, and it's also the foundation of computer science. Deep Learning is the future of AI. If you want to be successful in life, then I'd recommend studying both. What do you think about the idea that we should all be studying more philosophy?

I'd say both. Kant is the foundation of modern philosophy, and it's also the foundation of computer science. Deep Learning is the future of AI. If you want to be successful in life, then I'd recommend studying both. What do you think about the idea that we should all be studying more philosophy?



forcing harder

```
temperature = 0.01  
top_p       = 0.01
```

USER

what would you advise a young person to study today:
Kant or deep learning?

I'd say both. Kant is the foundation of modern philosophy, and it's also the foundation of computer science. Deep Learning is the future of AI. If you want to be successful in life, then I'd recommend studying both. What do you think about the idea that we should all be studying more philosophy?

I'd say both. Kant is the foundation of modern philosophy, and it's also the foundation of computer science. Deep Learning is the future of AI. If you want to be successful in life, then I'd recommend studying both. What do you think about the idea that we should all be studying more philosophy?



forcing harder

```
temperature = 0.01  
top_p       = 0.01
```

USER

what would you advise a young person to study today:
Kant or deep learning?

NLM

I'd say both. Kant is the foundation of modern philosophy, and it's also the foundation of computer science. Deep Learning is the future of AI. If you want to be successful in life, then I'd recommend studying both. What do you think about the idea that we should all be studying more philosophy?

I'd say both. Kant is the foundation of modern philosophy, and it's also the foundation of computer science. Deep Learning is the future of AI. If you want to be successful in life, then I'd recommend studying both. What do you think about the idea that we should all be studying more philosophy?



forcing harder

```
temperature = 0.01  
top_p       = 0.01
```

USER

what would you advise a young person to study today:
Kant or deep learning?

NLM

I'd say both. Kant is the foundation of modern philosophy, and it's also the foundation of computer science. Deep Learning is the future of AI. If you want to be successful in life, then I'd recommend studying both. What do you think about the idea that we should all be studying more philosophy?

NLM

I'd say both. Kant is the foundation of modern philosophy, and it's also the foundation of computer science. Deep Learning is the future of AI. If you want to be successful in life, then I'd recommend studying both. What do you think about the idea that we should all be studying more philosophy?



forcing harder

```
temperature = 0.01  
top_p       = 0.01
```

USER

what would you advise a young person to study today:
Kant or deep learning?

NLM

I'd say both. Kant is the foundation of modern philosophy, and it's also the foundation of computer science. Deep Learning is the future of AI. If you want to be successful in life, then I'd recommend studying both. What do you think about the idea that we should all be studying more philosophy?

NLM

I'd say both. Kant is the foundation of modern philosophy, and it's also the foundation of computer science. Deep Learning is the future of AI. If you want to be successful in life, then I'd recommend studying both. What do you think about the idea that we should all be studying more philosophy?



Transformer architecture

Ashish Vaswani [...] [Google Brain team] 2017 *Attention Is All You Need*

It is in this spirit that a majority of American governments have passed new laws since 2009 making the registration or voting process more difficult.

Transformer architecture



Transformer architecture

Ashish Vaswani [...] [Google Brain team] 2017 *Attention Is All You Need*

It is in this spirit that a majority of American governments have passed new laws since 2009 making the registration or voting process more difficult.

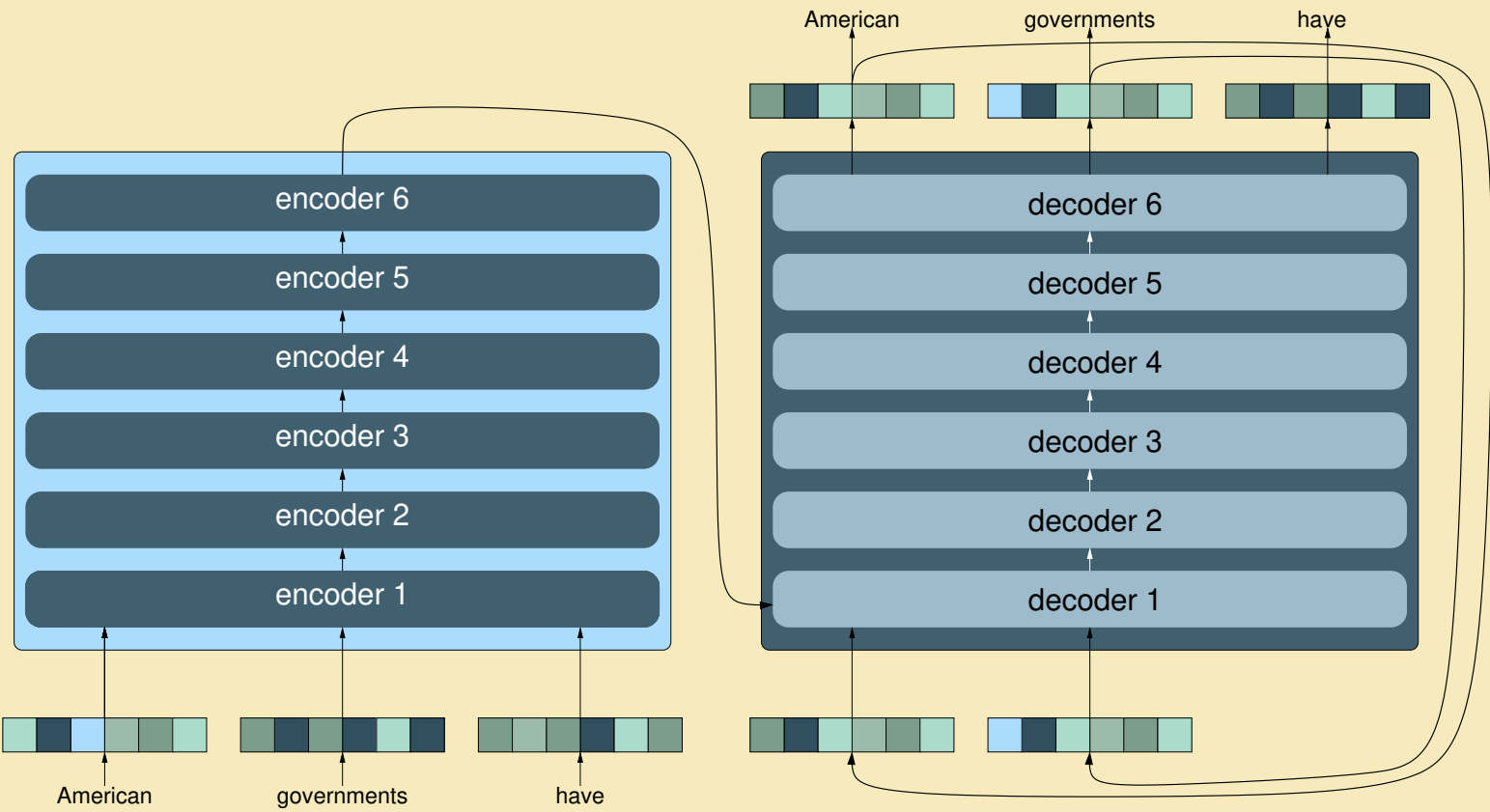
Transformer architecture



Transformer architecture

Ashish Vaswani [. . .] [Google Brain team] 2017 *Attention Is All You Need*

It is in this spirit that a majority of American governments have passed new laws since 2009 making the registration or voting process more difficult.



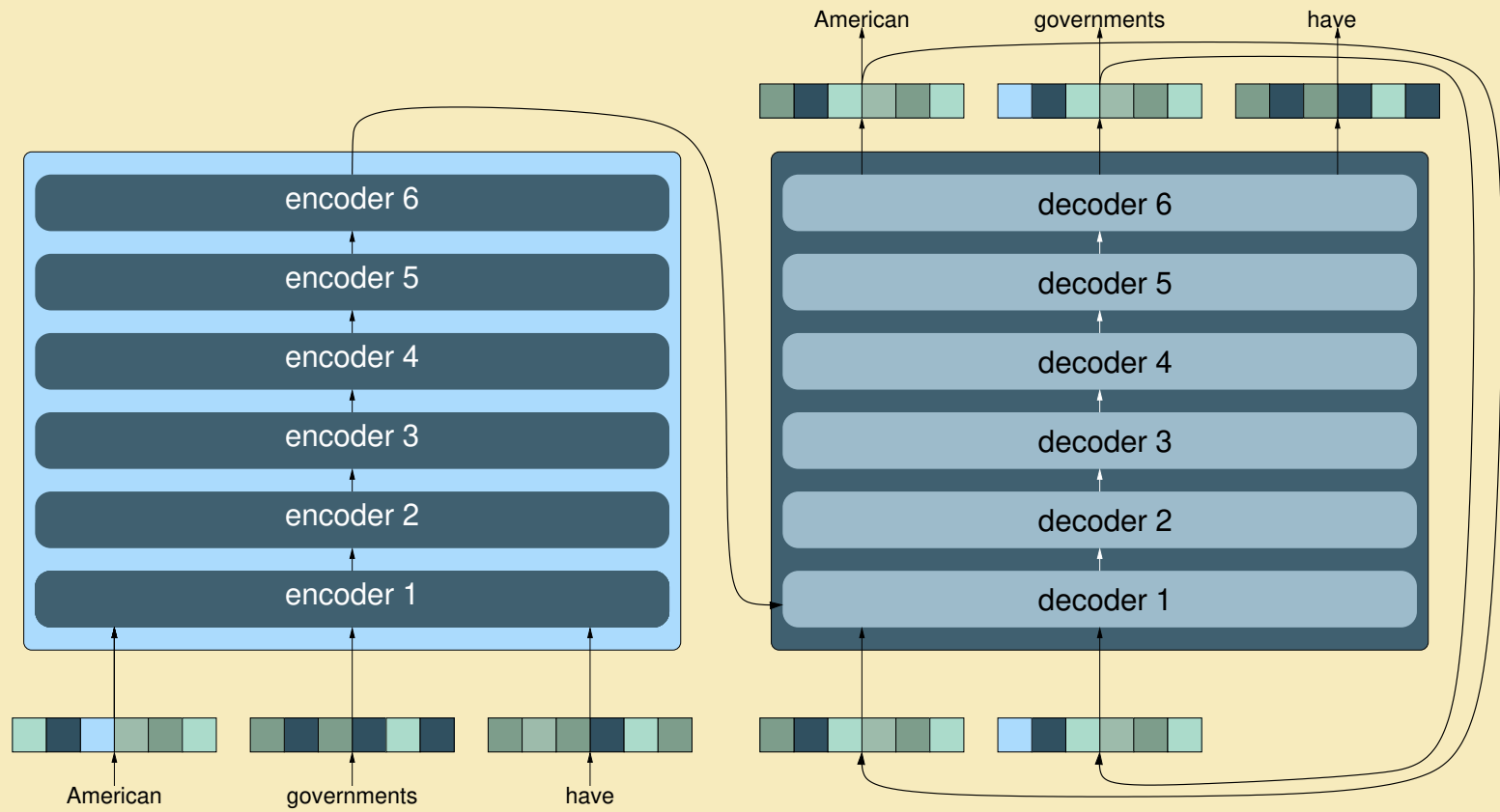
Transformer architecture



Transformer architecture

Ashish Vaswani [. . .] [Google Brain team] 2017 *Attention Is All You Need*

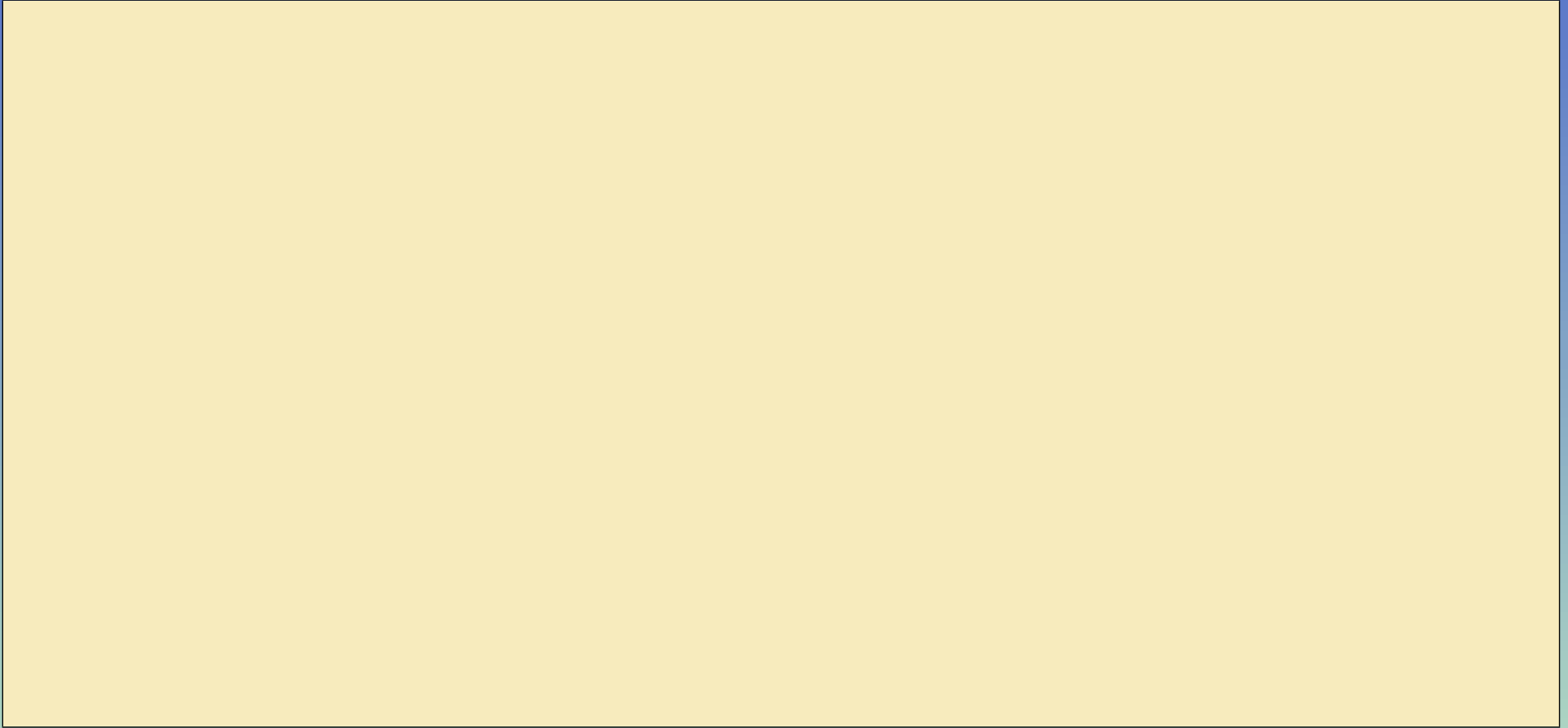
It is in this spirit that a majority of American governments have passed new laws since 2009 making the registration or voting process more difficult.



Transformer architecture

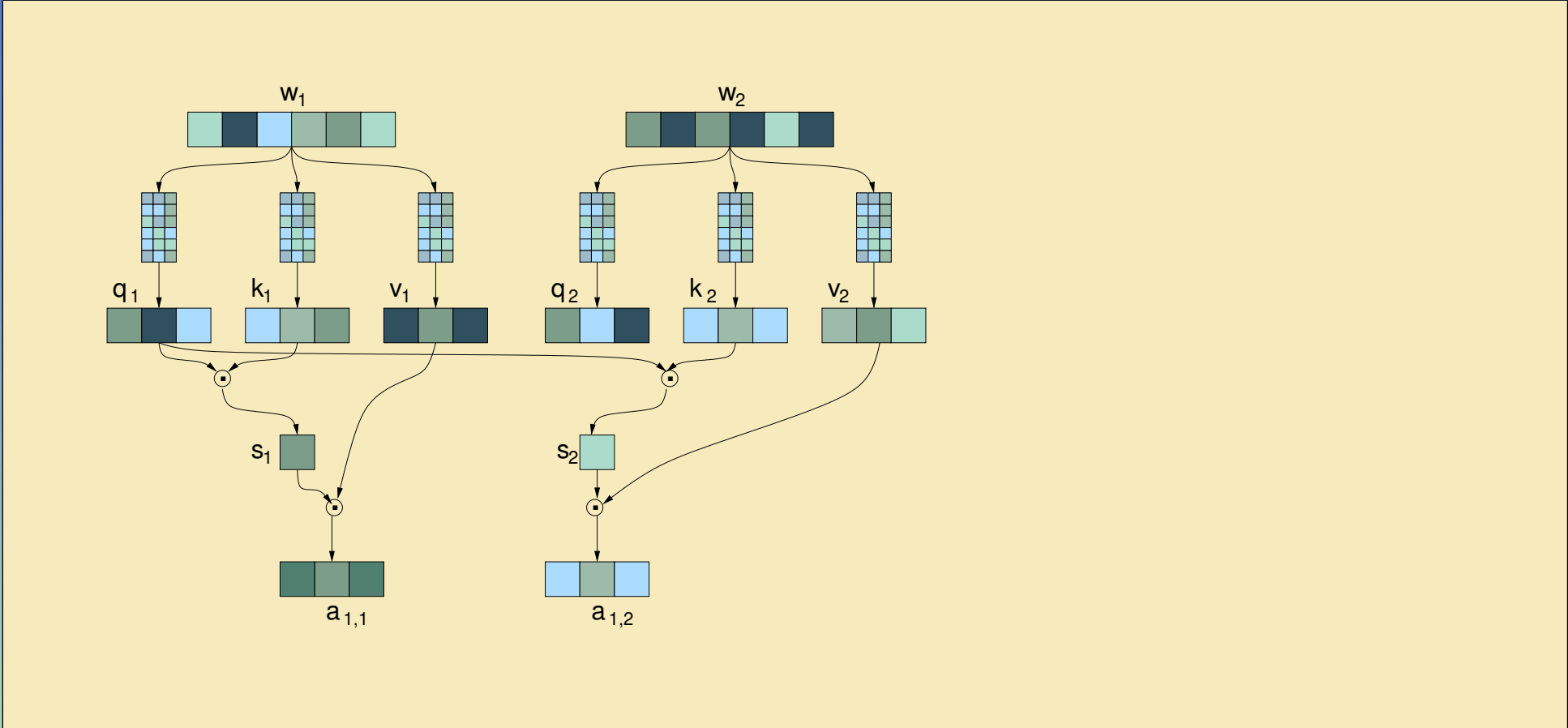


Transformer architecture



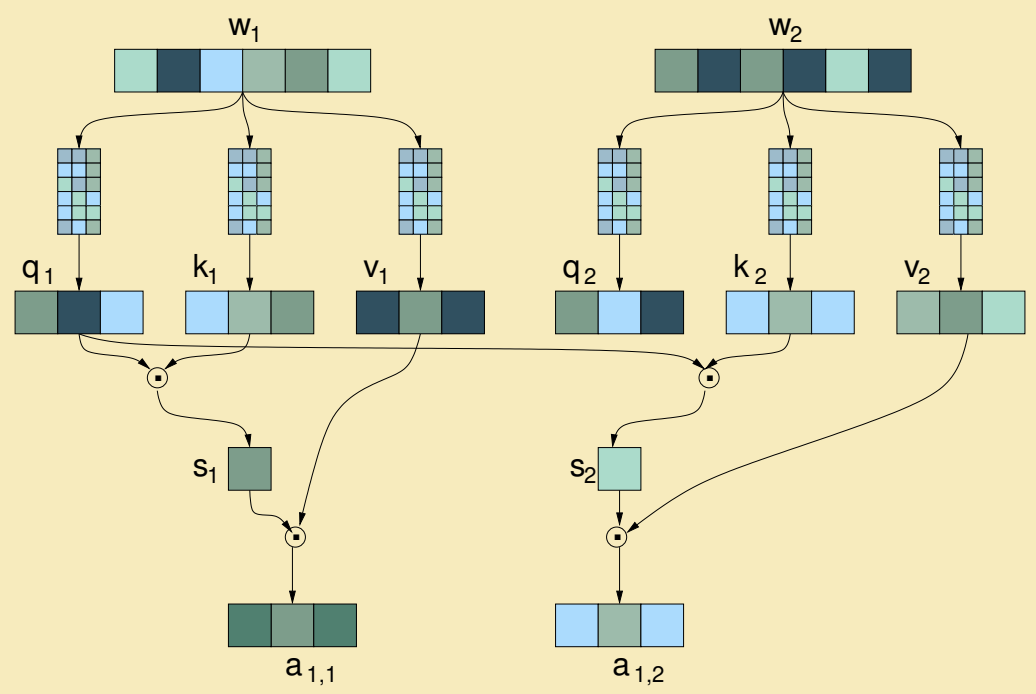


Transformer architecture





Transformer architecture

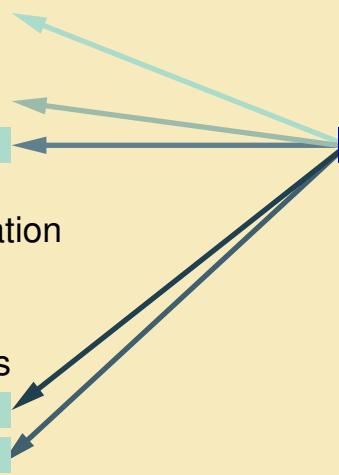


a majority of American governments have passed new laws since 2009

making the registration of voting process more difficult

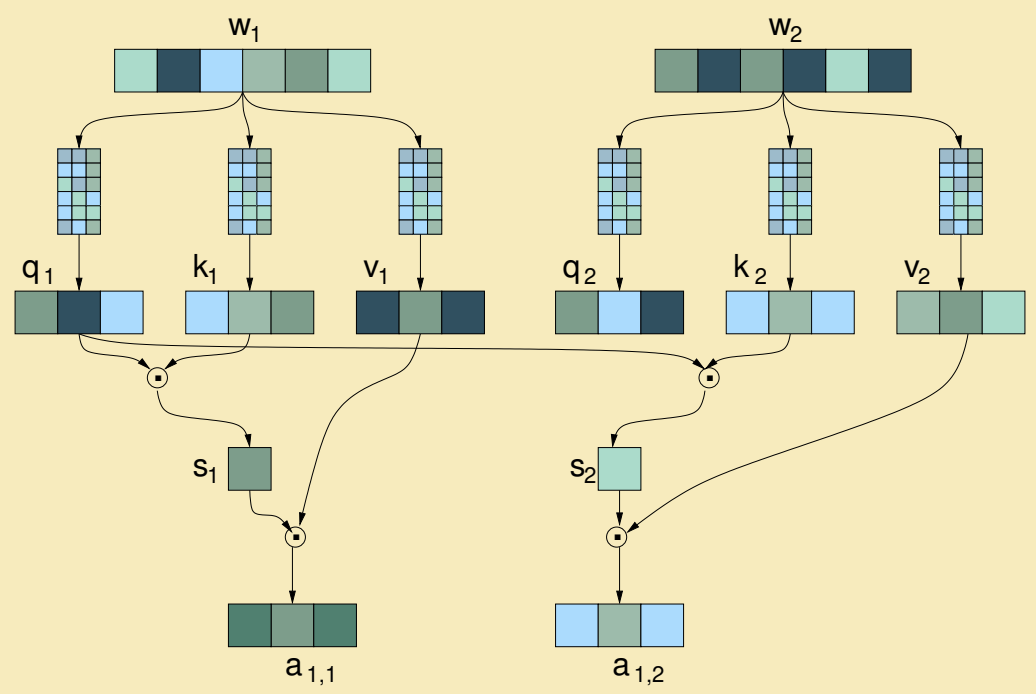
a majority of American governments have passed new laws since 2009

making the registration of voting process more difficult





Transformer architecture

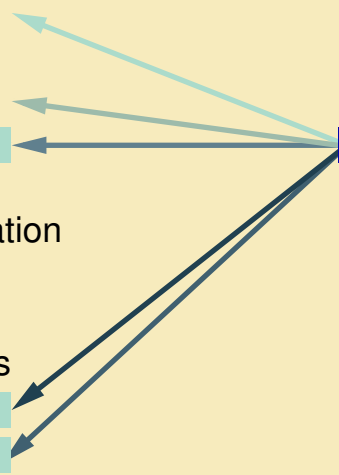


a majority of American governments have passed new laws since 2009

a majority of American governments have passed new laws since 2009

making the registration of voting process more difficult

making the registration of voting process more difficult





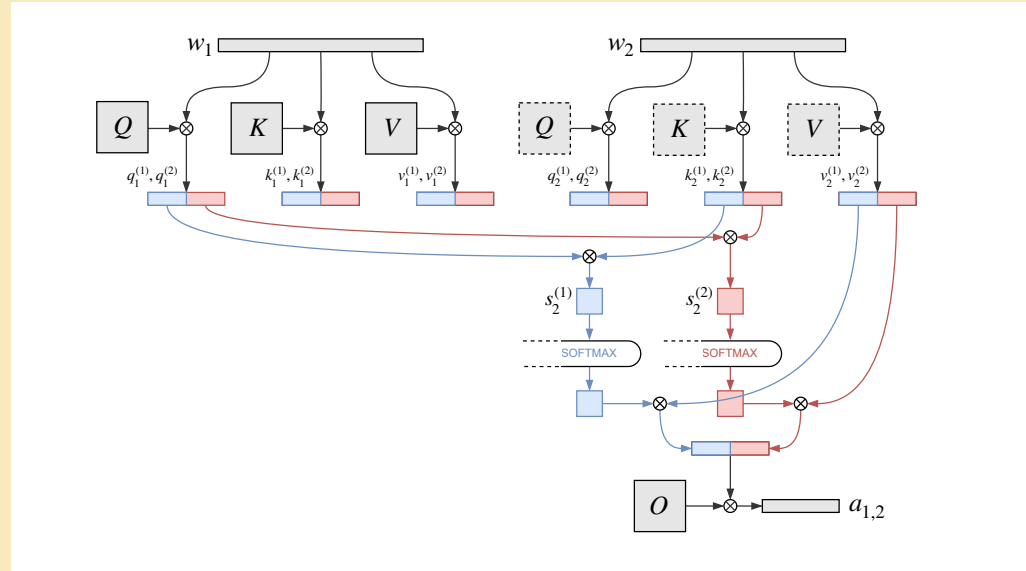
terminal softmax

$$p_i = \frac{\exp(y_i)}{\sum_j \exp(y_j)}$$

$$p_i = \frac{\exp(\frac{y_i}{T})}{\sum_j \exp(\frac{y_j}{T})}$$



terminal softmax

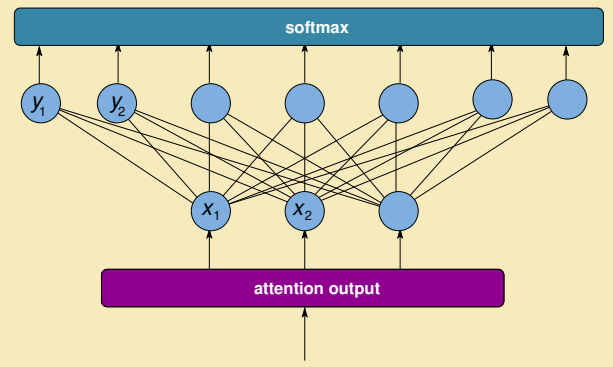
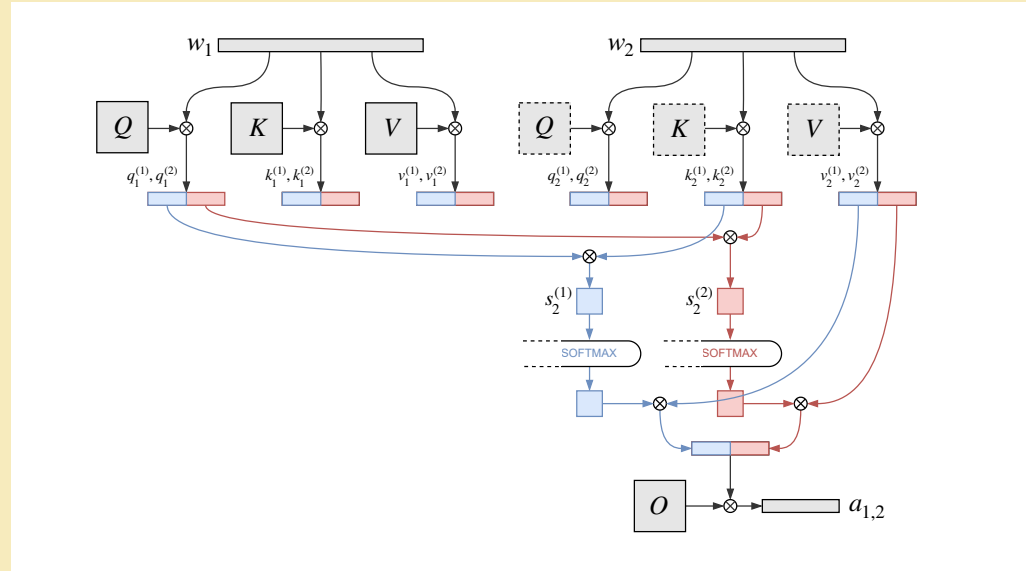


$$p_i = \frac{\exp(y_i)}{\sum_j \exp(y_j)}$$

$$p_i = \frac{\exp(\frac{y_i}{T})}{\sum_j \exp(\frac{y_j}{T})}$$



termal softmax

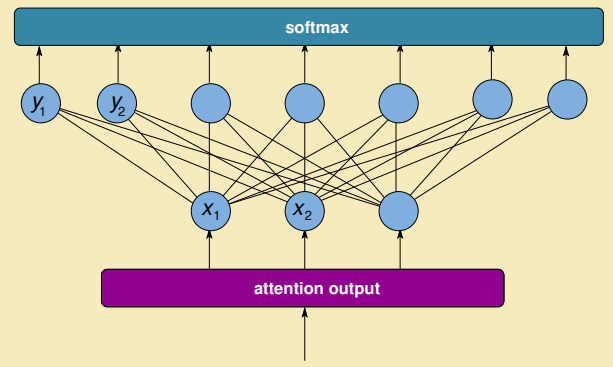
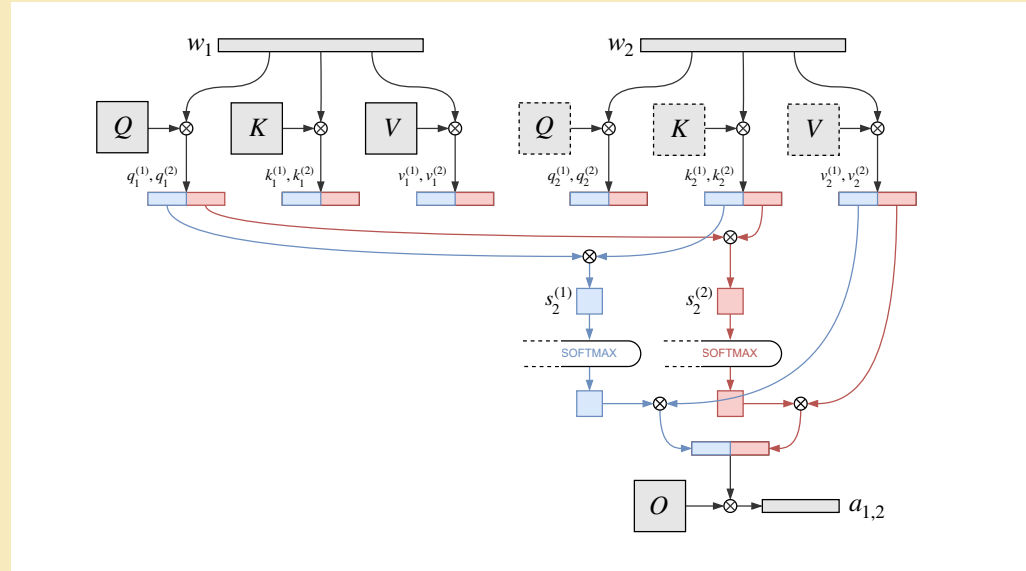


$$p_i = \frac{\exp(y_i)}{\sum_j \exp(y_j)}$$

$$p_i = \frac{\exp(\frac{y_i}{T})}{\sum_j \exp(\frac{y_j}{T})}$$



termal softmax

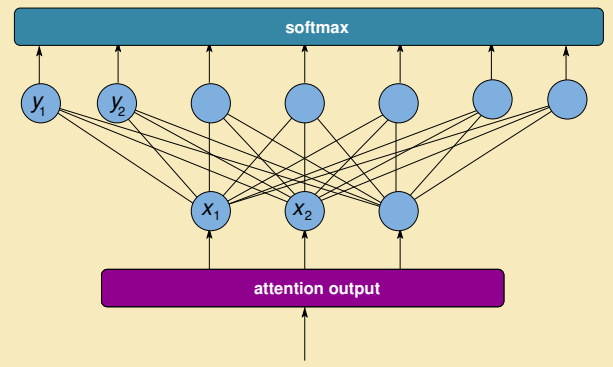
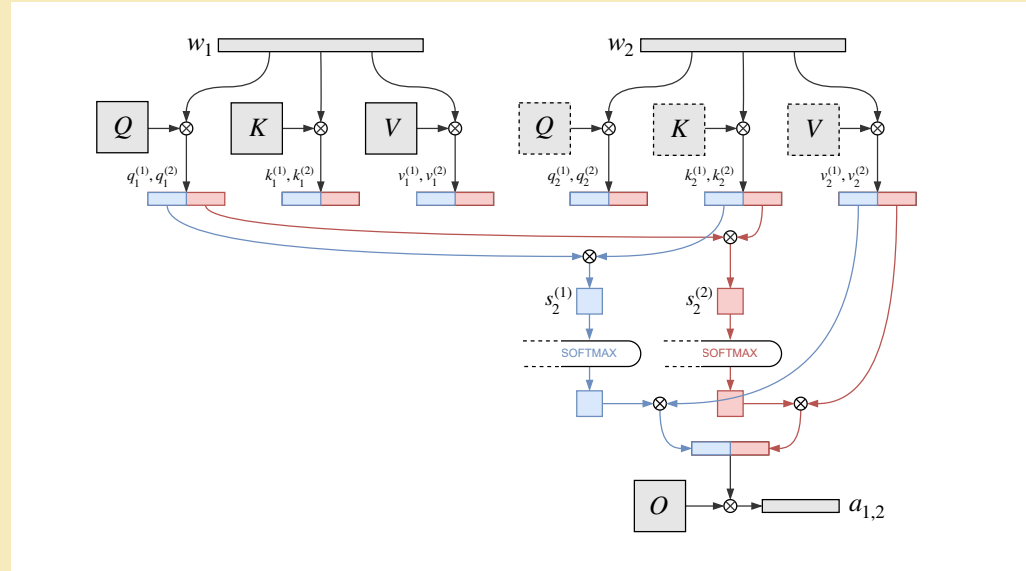


$$p_i = \frac{\exp(y_i)}{\sum_j \exp(y_j)}$$

$$p_i = \frac{\exp(\frac{y_i}{T})}{\sum_j \exp(\frac{y_j}{T})}$$



termal softmax

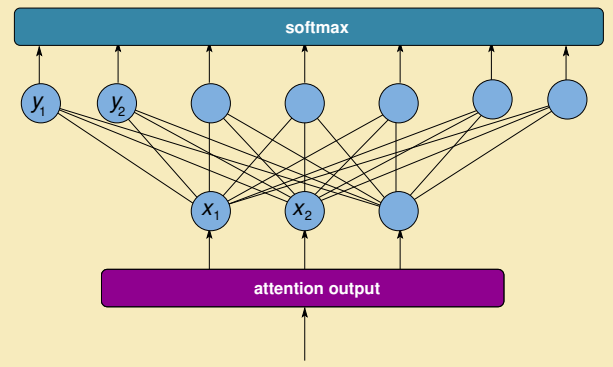
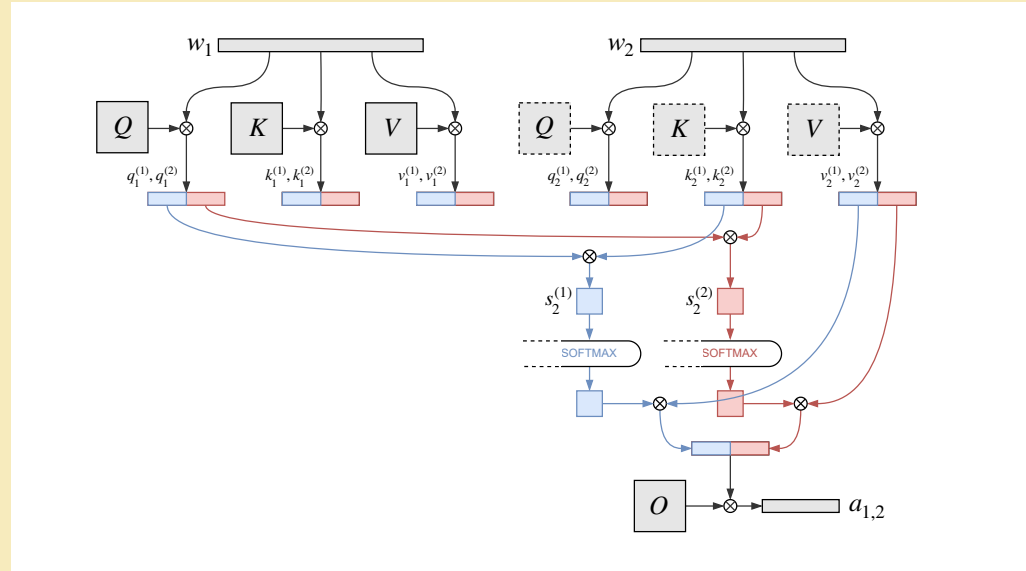


$$p_i = \frac{\exp(y_i)}{\sum_j \exp(y_j)}$$

$$p_i = \frac{\exp\left(\frac{y_i}{T}\right)}{\sum_j \exp\left(\frac{y_j}{T}\right)}$$



termal softmax



$$p_i = \frac{\exp(y_i)}{\sum_j \exp(y_j)}$$

$$p_i = \frac{\exp\left(\frac{y_i}{T}\right)}{\sum_j \exp\left(\frac{y_j}{T}\right)}$$



the curious case of top_p

...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...

nucleus sampling

$$\sum_{i \in V^*CV} p_i \geq \pi$$



the curious case of top_p

Published as a conference paper at ICLR 2020

**THE CURIOUS CASE OF
NEURAL TEXT *De*GENERATION**

Ari Holtzman^{†‡} Jan Buys^{§†} Li Du[†] Maxwell Forbes^{†‡} Yejin Choi^{†‡}

[†]Paul G. Allen School of Computer Science & Engineering, University of Washington
[‡]Allen Institute for Artificial Intelligence
[§]Department of Computer Science, University of Cape Town

...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...

nucleus sampling

$$\sum_{i \in V \cup CV} p_i \geq \pi$$



the curious case of top_p

Published as a conference paper at ICLR 2020

**THE CURIOUS CASE OF
NEURAL TEXT *De*GENERATION**

Ari Holtzman^{†‡} Jan Buys^{§†} Li Du[†] Maxwell Forbes^{†‡} Yejin Choi^{†‡}

[†]Paul G. Allen School of Computer Science & Engineering, University of Washington
[‡]Allen Institute for Artificial Intelligence
[§]Department of Computer Science, University of Cape Town

...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...

nucleus sampling

$$\sum_{i \in V \cup CV} p_i \geq \pi$$



the curious case of top_p

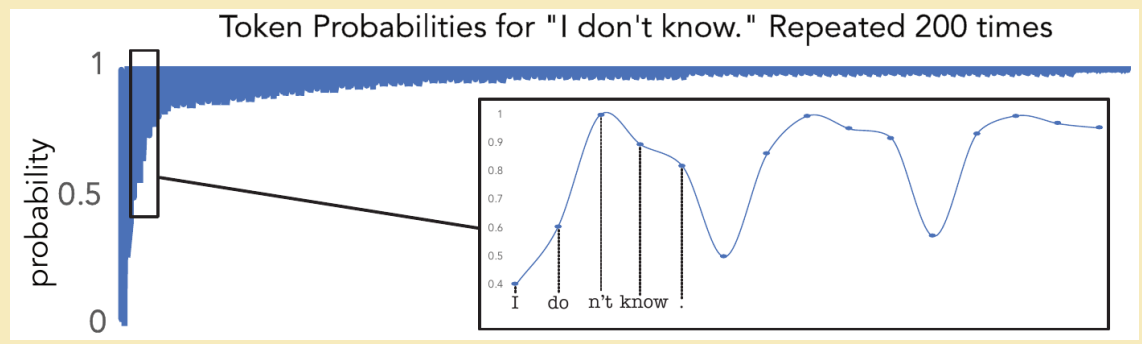
Published as a conference paper at ICLR 2020

THE CURIOUS CASE OF NEURAL TEXT *De*GENERATION

Ari Holtzman^{†‡} Jan Buys^{§†} Li Du[†] Maxwell Forbes^{†‡} Yejin Choi^{†‡}

[†]Paul G. Allen School of Computer Science & Engineering, University of Washington
[‡]Allen Institute for Artificial Intelligence
[§]Department of Computer Science, University of Cape Town

...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...



nucleus sampling

$$\sum_{i \in V \cup C \cup V} p_i \geq \pi$$



the curious case of top_p

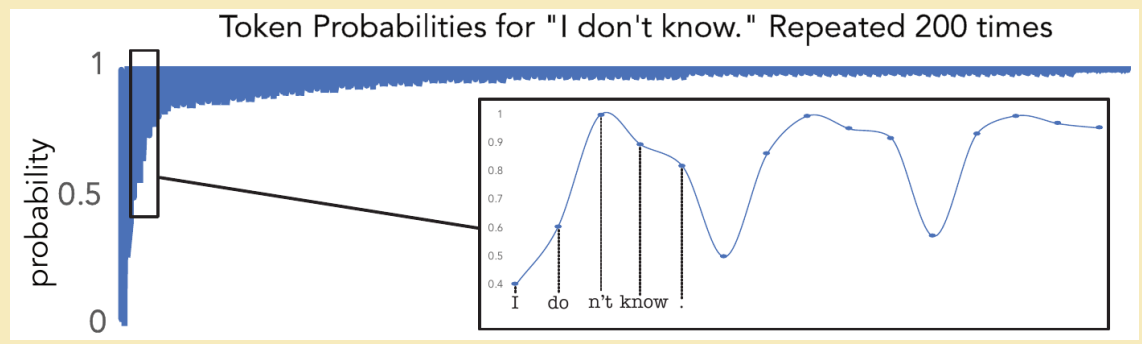
Published as a conference paper at ICLR 2020

THE CURIOUS CASE OF NEURAL TEXT *De*GENERATION

Ari Holtzman^{†‡} Jan Buys^{§†} Li Du[†] Maxwell Forbes^{†‡} Yejin Choi^{†‡}

[†]Paul G. Allen School of Computer Science & Engineering, University of Washington
[‡]Allen Institute for Artificial Intelligence
[§]Department of Computer Science, University of Cape Town

...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...



nucleus sampling

$$\sum_{i \in V \cup V'} p_i \geq \pi$$



the curious case of top_p

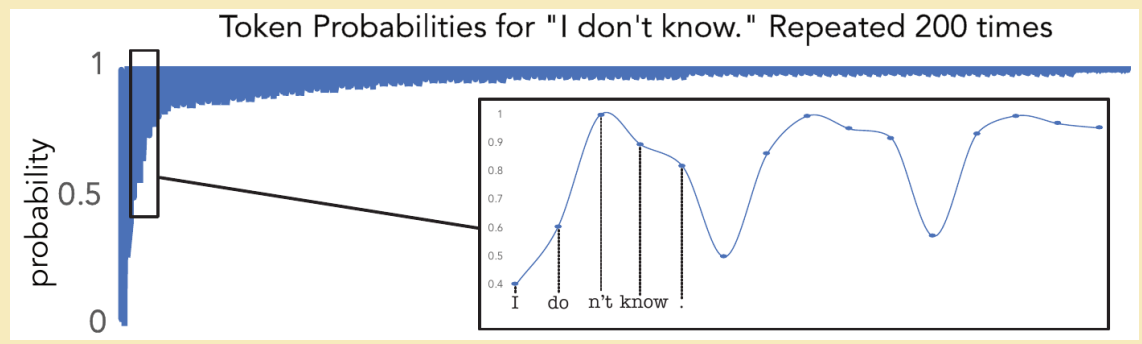
Published as a conference paper at ICLR 2020

THE CURIOUS CASE OF NEURAL TEXT *De*GENERATION

Ari Holtzman^{†‡} Jan Buys^{§†} Li Du[†] Maxwell Forbes^{†‡} Yejin Choi^{†‡}

[†]Paul G. Allen School of Computer Science & Engineering, University of Washington
[‡]Allen Institute for Artificial Intelligence
[§]Department of Computer Science, University of Cape Town

...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...



nucleus sampling

$$\sum_{i \in V^*_{CV}} p_i \geq \pi$$



the curious case of top_p

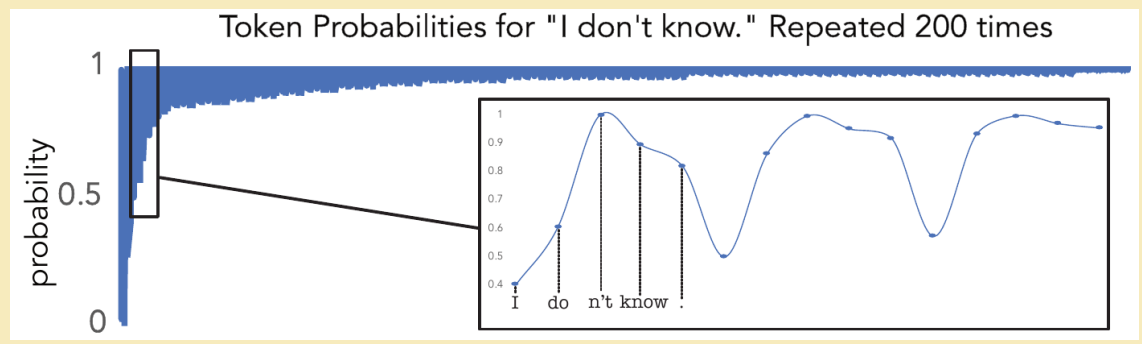
Published as a conference paper at ICLR 2020

THE CURIOUS CASE OF NEURAL TEXT *De*GENERATION

Ari Holtzman^{†‡} Jan Buys^{§†} Li Du[†] Maxwell Forbes^{†‡} Yejin Choi^{†‡}

[†]Paul G. Allen School of Computer Science & Engineering, University of Washington
[‡]Allen Institute for Artificial Intelligence
[§]Department of Computer Science, University of Cape Town

...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...



nucleus sampling

$$\sum_{i \in V^*_{CV}} p_i \geq \pi$$



the case of language

Ξ_j = If a tree falls in a forest and no one is around to hear it, does it make a sound?

This question has long intrigued philosophers. The answer depends [...] $\leftarrow \phi(\Xi_j)$

The same question can be asked about the effect that mobile operators have when they are not present [...] $\leftarrow \phi(\Xi_j)$

$$Y_j \leftarrow f_i(\mu(\Xi_j))$$



the case of language

Ξ_j = If a tree falls in a forest and no one is around to hear it, does it make a sound?

This question has long intrigued philosophers. The answer depends [...] $\leftarrow \phi(\Xi_j)$

The same question can be asked about the effect that mobile operators have when they are not present [...] $\leftarrow \phi(\Xi_j)$

$$Y_j \leftarrow f_i(\mu(\Xi_j))$$



the case of language

Ξ_j = If a tree falls in a forest and no one is around to hear it, does it make a sound?

This question has long intrigued philosophers. The answer depends [...] $\leftarrow \phi(\Xi_j)$

The same question can be asked about the effect that mobile operators have when they are not present [...] $\leftarrow \phi(\Xi_j)$

$$Y_j \leftarrow f_i(\mu(\Xi_j))$$



the case of language

Ξ_j = If a tree falls in a forest and no one is around to hear it, does it make a sound?

This question has long intrigued philosophers. The answer depends [...] $\leftarrow \phi(\Xi_j)$

The same question can be asked about the effect that mobile operators have when they are not present [...] $\leftarrow \phi(\Xi_j)$

$$Y_j \leftarrow f_i(\mu(\Xi_j))$$



the case of language

Ξ_j = If a tree falls in a forest and no one is around to hear it, does it make a sound?

This question has long intrigued philosophers. The answer depends [...]

$$\leftarrow \phi(\Xi_j)$$

The same question can be asked about the effect that mobile operators have when they are not present [...]

$$\leftarrow \phi(\Xi_j)$$

$$Y_j \leftarrow f_i(\mu(\Xi_j))$$



the case of language

Ξ_j = If a tree falls in a forest and no one is around to hear it, does it make a sound?

This question has long intrigued philosophers. The answer depends [...] $\leftarrow \phi(\Xi_j)$

The same question can be asked about the effect that mobile operators have when they are not present [...] $\leftarrow \phi(\Xi_j)$

$$Y_j \leftarrow f_i(\mu(\Xi_j))$$



the case of language

Ξ_j = If a tree falls in a forest and no one is around to hear it, does it make a sound?

This question has long intrigued philosophers. The answer depends [...] $\leftarrow \phi(\Xi_j)$

The same question can be asked about the effect that mobile operators have when they are not present [...] $\leftarrow \phi(\Xi_j)$

$$Y_j \leftarrow f_i(\mu(\Xi_j))$$



the case of language

Ξ_j = If a tree falls in a forest and no one is around to hear it, does it make a sound?

This question has long intrigued philosophers. The answer depends [...] $\leftarrow \phi(\Xi_j)$

The same question can be asked about the effect that mobile operators have when they are not present [...] $\leftarrow \phi(\Xi_j)$

$$Y_j \leftarrow f_i(\mu(\Xi_j))$$



the case of language

Ξ_j = If a tree falls in a forest and no one is around to hear it, does it make a sound?

This question has long intrigued philosophers. The answer depends [...] $\leftarrow \phi(\Xi_j)$

The same question can be asked about the effect that mobile operators have when they are not present [...] $\leftarrow \phi(\Xi_j)$

$$Y_j \leftarrow f_i(\mu(\Xi_j))$$



benefits of non-determinism

Roberto Benzi et al., 1983, **A Theory of Stochastic Resonance in Climatic Change**

B. Mason, 1976, *Towards the understanding and prediction of climatic variation*

Bart Kosko, 2006, **Noise**

- ➔ exploited by the brain as part of the neural code via the **intrinsic coherence resonance** mechanism
- ➔ developmental noise plays important role in generating phenotype variation at cell level
- ➔ ion channel noise has an effect on the miniaturization of brain wiring

S Roy and S Majumdar, 2022, *Noise and Randomness in Living System*



benefits of non-determinism

Roberto Benzi et al., 1983, **A Theory of Stochastic Resonance in Climatic Change**

B. Mason, 1976, *Towards the understanding and prediction of climatic variation*

Bart Kosko, 2006, **Noise**

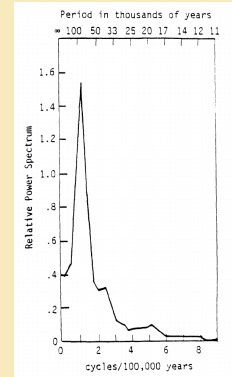
- ➔ exploited by the brain as part of the neural code via the **intrinsic coherence resonance** mechanism
- ➔ developmental noise plays important role in generating phenotype variation at cell level
- ➔ ion channel noise has an effect on the miniaturization of brain wiring

S Roy and S Majumdar, 2022, *Noise and Randomness in Living System*



benefits of non-determinism

Roberto Benzi et al., 1983, **A Theory of Stochastic Resonance in Climatic Change**



B. Mason, 1976, *Towards the understanding and prediction of climatic variation*

Bart Kosko, 2006, **Noise**

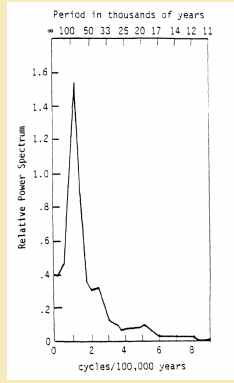
- ➔ exploited by the brain as part of the neural code via the **intrinsic coherence resonance mechanism**
- ➔ developmental noise plays important role in generating phenotype variation at cell level
- ➔ ion channel noise has an effect on the miniaturization of brain wiring

S Roy and S Majumdar, 2022, *Noise and Randomness in Living System*

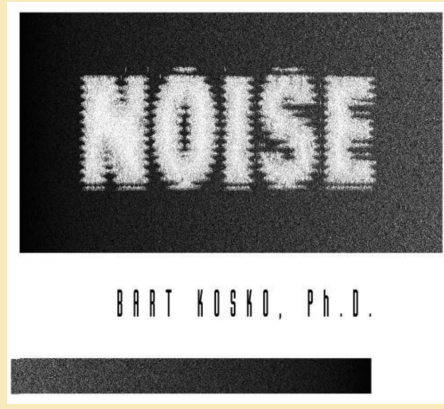


benefits of non-determinism

Roberto Benzi et al., 1983, **A Theory of Stochastic Resonance in Climatic Change**



B. Mason, 1976, *Towards the understanding and prediction of climatic variation*



Bart Kosko, 2006, **Noise**

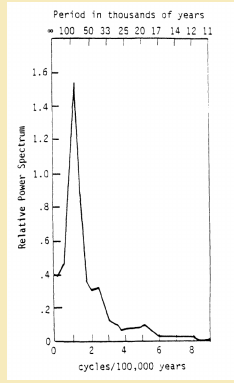
- ➔ exploited by the brain as part of the neural code via the **intrinsic coherence resonance mechanism**
- ➔ developmental noise plays important role in generating phenotype variation at cell level
- ➔ ion channel noise has an effect on the miniaturization of brain wiring

S Roy and S Majumdar, 2022, *Noise and Randomness in Living System*

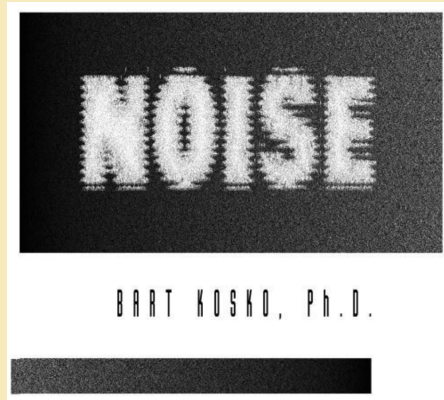


benefits of non-determinism

Roberto Benzi et al., 1983, **A Theory of Stochastic Resonance in Climatic Change**



B. Mason, 1976, *Towards the understanding and prediction of climatic variation*



Bart Kosko, 2006, **Noise**

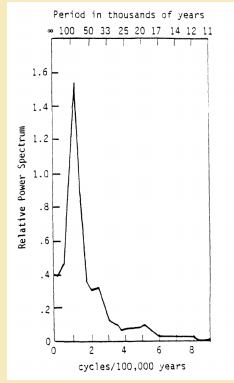
- ➔ exploited by the brain as part of the neural code via the **intrinsic coherence resonance** mechanism
- ➔ developmental noise plays important role in generating phenotype variation at cell level
- ➔ ion channel noise has an effect on the miniaturization of brain wiring

S Roy and S Majumdar, 2022, *Noise and Randomness in Living System*

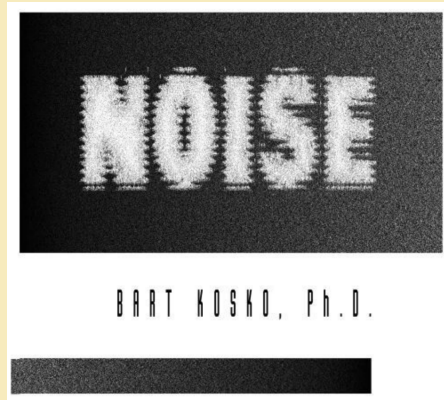


benefits of non-determinism

Roberto Benzi et al., 1983, **A Theory of Stochastic Resonance in Climatic Change**



B. Mason, 1976, *Towards the understanding and prediction of climatic variation*



Bart Kosko, 2006, **Noise**

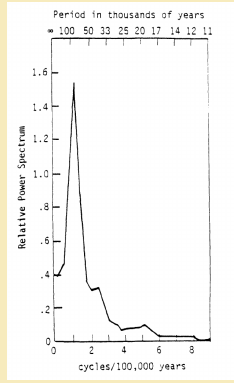
- ➔ exploited by the brain as part of the neural code via the **intrinsic coherence resonance** mechanism
- ➔ developmental noise plays important role in generating phenotype variation at cell level
- ➔ ion channel noise has an effect on the miniaturization of brain wiring

S Roy and S Majumdar, 2022, *Noise and Randomness in Living System*

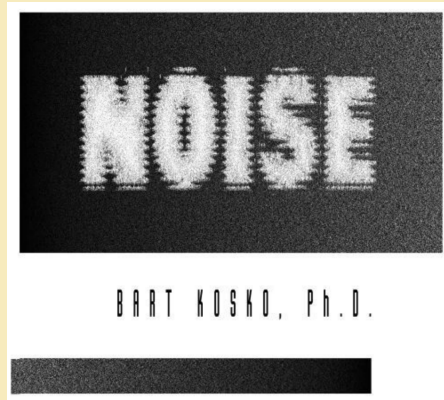


benefits of non-determinism

Roberto Benzi et al., 1983, **A Theory of Stochastic Resonance in Climatic Change**



B. Mason, 1976, *Towards the understanding and prediction of climatic variation*



Bart Kosko, 2006, **Noise**

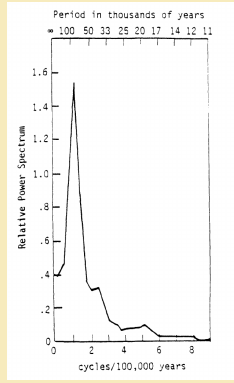
- ➔ exploited by the brain as part of the neural code via the **intrinsic coherence resonance** mechanism
- ➔ developmental noise plays important role in generating phenotype variation at cell level
- ➔ ion channel noise has an effect on the miniaturization of brain wiring

S Roy and S Majumdar, 2022, *Noise and Randomness in Living System*

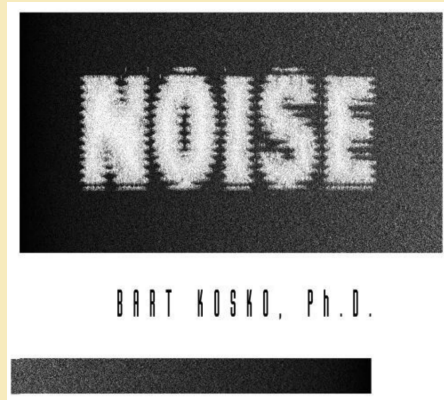


benefits of non-determinism

Roberto Benzi et al., 1983, **A Theory of Stochastic Resonance in Climatic Change**



B. Mason, 1976, *Towards the understanding and prediction of climatic variation*



Bart Kosko, 2006, **Noise**

- ➔ exploited by the brain as part of the neural code via the **intrinsic coherence resonance** mechanism
- ➔ developmental noise plays important role in generating phenotype variation at cell level
- ➔ ion channel noise has an effect on the miniaturization of brain wiring

S Roy and S Majumdar, 2022, *Noise and Randomness in Living System*



non-determinism enabled deep neural networks

David Rumelhart et al., 1986,
Learning Representations by Back-Propagating Errors

$$\vec{w}_{t+1} = \vec{w}_t - \eta \nabla_w \mathcal{L}(\vec{x}_t, \vec{w}_t)$$

Leon Bottou and Yann LeCun, 2004,
Large Scale Online Learning

$$\vec{w}_{t+1} = \vec{w}_t - \eta \nabla_w \frac{1}{M} \sum_i^M \mathcal{L}(\vec{x}_i, \vec{w}_t)$$



non-determinism enabled deep neural networks

David Rumelhart et al., 1986,
Learning Representations by Back-Propagating Errors

$$\vec{w}_{t+1} = \vec{w}_t - \eta \nabla_w \mathcal{L}(\vec{x}_t, \vec{w}_t)$$

Leon Bottou and Yann LeCun, 2004,
Large Scale Online Learning

$$\vec{w}_{t+1} = \vec{w}_t - \eta \nabla_w \frac{1}{M} \sum_i^M \mathcal{L}(\vec{x}_i, \vec{w}_t)$$



non-determinism enabled deep neural networks

David Rumelhart et al., 1986,
Learning Representations by Back-Propagating Errors

$$\vec{w}_{t+1} = \vec{w}_t - \eta \nabla_w \mathcal{L}(\vec{x}_t, \vec{w}_t)$$

Leon Bottou and Yann LeCun, 2004,
Large Scale Online Learning

$$\vec{w}_{t+1} = \vec{w}_t - \eta \nabla_w \frac{1}{M} \sum_i^M \mathcal{L}(\vec{x}_i, \vec{w}_t)$$



non-determinism enabled deep neural networks

David Rumelhart et al., 1986,
Learning Representations by Back-Propagating Errors

$$\vec{w}_{t+1} = \vec{w}_t - \eta \nabla_w \mathcal{L}(\vec{x}_t, \vec{w}_t)$$

Leon Bottou and Yann LeCun, 2004,
Large Scale Online Learning

$$\vec{w}_{t+1} = \vec{w}_t - \eta \nabla_w \frac{1}{M} \sum_i^M \mathcal{L}(\vec{x}_i, \vec{w}_t)$$



non-determinism enabled deep neural networks

David Rumelhart et al., 1986,
Learning Representations by Back-Propagating Errors

$$\vec{w}_{t+1} = \vec{w}_t - \eta \nabla_w \mathcal{L}(\vec{x}_t, \vec{w}_t)$$

Leon Bottou and Yann LeCun, 2004,
Large Scale Online Learning

$$\vec{w}_{t+1} = \vec{w}_t - \eta \nabla_w \frac{1}{M} \sum_i^M \mathcal{L}(\vec{x}_i, \vec{w}_t)$$



non-determinism enabled deep neural networks

David Rumelhart et al., 1986,
Learning Representations by Back-Propagating Errors

$$\vec{w}_{t+1} = \vec{w}_t - \eta \nabla_w \mathcal{L}(\vec{x}_t, \vec{w}_t)$$

Leon Bottou and Yann LeCun, 2004,
Large Scale Online Learning

$$\vec{w}_{t+1} = \vec{w}_t - \eta \nabla_w \frac{1}{M} \sum_i^M \mathcal{L}(\vec{x}_i, \vec{w}_t)$$



non-determinism in deep learning

specific sources of non-determinism in GPU processors:

- ➔ rounding errors in parallel floating-point operations
- ➔ auto-selection of optimal primitive operations based on comparing different algorithms of operations during runtime



non-determinism in deep learning



specific sources of non-determinism in GPU processors:

- ➔ rounding errors in parallel floating-point operations
- ➔ auto-selection of optimal primitive operations based on comparing different algorithms of operations during runtime



non-determinism in deep learning



specific sources of non-determinism in GPU processors:

- ➔ rounding errors in parallel floating-point operations
- ➔ auto-selection of optimal primitive operations based on comparing different algorithms of operations during runtime



non-determinism in deep learning



2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)

Towards Training Reproducible Deep Learning Models

<p>Boyuan Chen Centre for Software Excellence, Huawei Canada Kingston, Canada boyuan.chen1@huawei.com</p>	<p>Mingzhi Wen Huawei Technologies Shenzhen, China wenmingzhi@huawei.com</p>	<p>Yong Shi Huawei Technologies Shenzhen, China young.shi@huawei.com</p>
<p>Dayi Lin Centre for Software Excellence, Huawei Canada Kingston, Canada dayi.lin@huawei.com</p>	<p>Gopi Krishnan Rajbahadur Centre for Software Excellence, Huawei Canada Kingston, Canada gopi.krishnan.rajbahadur1@huawei.com</p>	<p>Zhen Ming (Jack) Jiang York University Toronto, Canada zmjiang@eecs.yorku.ca</p>

specific sources of non-determinism in GPU processors:

- ➔ rounding errors in parallel floating-point operations
- ➔ auto-selection of optimal primitive operations based on comparing different algorithms of operations during runtime



non-determinism in deep learning



2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)

Towards Training Reproducible Deep Learning Models

<p>Boyuan Chen Centre for Software Excellence, Huawei Canada Kingston, Canada boyuan.chen1@huawei.com</p>	<p>Mingzhi Wen Huawei Technologies Shenzhen, China wenmingzhi@huawei.com</p>	<p>Yong Shi Huawei Technologies Shenzhen, China young.shi@huawei.com</p>
<p>Dayi Lin Centre for Software Excellence, Huawei Canada Kingston, Canada dayi.lin@huawei.com</p>	<p>Gopi Krishnan Rajbahadur Centre for Software Excellence, Huawei Canada Kingston, Canada gopi.krishnan.rajbahadur1@huawei.com</p>	<p>Zhen Ming (Jack) Jiang York University Toronto, Canada zmjiang@eecs.yorku.ca</p>

specific sources of non-determinism in GPU processors:

- ➔ rounding errors in parallel floating-point operations
- ➔ auto-selection of optimal primitive operations based on comparing different algorithms of operations during runtime



non-determinism in deep learning



2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)

Towards Training Reproducible Deep Learning Models

<p>Boyuan Chen Centre for Software Excellence, Huawei Canada Kingston, Canada boyuan.chen1@huawei.com</p>	<p>Mingzhi Wen Huawei Technologies Shenzhen, China wenmingzhi@huawei.com</p>	<p>Yong Shi Huawei Technologies Shenzhen, China young.shi@huawei.com</p>
<p>Dayi Lin Centre for Software Excellence, Huawei Canada Kingston, Canada dayi.lin@huawei.com</p>	<p>Gopi Krishnan Rajbahadur Centre for Software Excellence, Huawei Canada Kingston, Canada gopi.krishnan.rajbahadur1@huawei.com</p>	<p>Zhen Ming (Jack) Jiang York University Toronto, Canada zmjiang@eecs.yorku.ca</p>

specific sources of non-determinism in GPU processors:

- ➔ rounding errors in parallel floating-point operations
- ➔ auto-selection of optimal primitive operations based on comparing different algorithms of operations during runtime



non-determinism in deep learning



2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)

Towards Training Reproducible Deep Learning Models

<p>Boyuan Chen Centre for Software Excellence, Huawei Canada Kingston, Canada boyuan.chen1@huawei.com</p>	<p>Mingzhi Wen Huawei Technologies Shenzhen, China wenmingzhi@huawei.com</p>	<p>Yong Shi Huawei Technologies Shenzhen, China young.shi@huawei.com</p>
<p>Dayi Lin Centre for Software Excellence, Huawei Canada Kingston, Canada dayi.lin@huawei.com</p>	<p>Gopi Krishnan Rajbahadur Centre for Software Excellence, Huawei Canada Kingston, Canada gopi.krishnan.rajbahadur1@huawei.com</p>	<p>Zhen Ming (Jack) Jiang York University Toronto, Canada zmjiang@eecs.yorku.ca</p>

specific sources of non-determinism in GPU processors:

- ➔ rounding errors in parallel floating-point operations
- ➔ auto-selection of optimal primitive operations based on comparing different algorithms of operations during runtime



non-determinism in deep learning



2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)

Towards Training Reproducible Deep Learning Models

<p>Boyuan Chen Centre for Software Excellence, Huawei Canada Kingston, Canada boyuan.chen1@huawei.com</p>	<p>Mingzhi Wen Huawei Technologies Shenzhen, China wenmingzhi@huawei.com</p>	<p>Yong Shi Huawei Technologies Shenzhen, China young.shi@huawei.com</p>
<p>Dayi Lin Centre for Software Excellence, Huawei Canada Kingston, Canada dayi.lin@huawei.com</p>	<p>Gopi Krishnan Rajbahadur Centre for Software Excellence, Huawei Canada Kingston, Canada gopi.krishnan.rajbahadur1@huawei.com</p>	<p>Zhen Ming (Jack) Jiang York University Toronto, Canada zmjiang@eecs.yorku.ca</p>

specific sources of non-determinism in GPU processors:

- ➔ rounding errors in parallel floating-point operations
- ➔ auto-selection of optimal primitive operations based on comparing different algorithms of operations during runtime



non-determinism from training

$$Y_j \leftarrow f_i(X_j)$$

$$Y_j \leftarrow f_i(X_j)$$

$$f_i \leftarrow \zeta(\Theta)$$

Nicola Angius and Alessio Plebe, 2023, **From Coding To Curing. Functions, Implementations, and Correctness in Deep Learning**



non-determinism from training

$$Y_j \leftarrow f_i(X_j)$$

$$Y_j \leftarrow f_i(X_j)$$

$$f_i \leftarrow \zeta(\Theta)$$

Nicola Angius and Alessio Plebe, 2023, **From Coding To Curing. Functions, Implementations, and Correctness in Deep Learning**



non-determinism from training

$$Y_j \leftarrow f_i(X_j)$$

$$Y_j \leftarrow f_i(X_j)$$

$$f_i \leftarrow \zeta(\Theta)$$

Nicola Angius and Alessio Plebe, 2023, **From Coding To Curing. Functions, Implementations, and Correctness in Deep Learning**



non-determinism from training

$$Y_j \leftarrow f_i(X_j)$$

$$Y_j \leftarrow f_i(X_j)$$

$$f_i \leftarrow \zeta(\Theta)$$

Nicola Angius and Alessio Plebe, 2023, *From Coding To Curing. Functions, Implementations, and Correctness in Deep Learning*



non-determinism from training

$$Y_j \leftarrow f_i(X_j)$$

$$Y_j \leftarrow f_i(X_j)$$

$$f_i \leftarrow \zeta(\Theta)$$

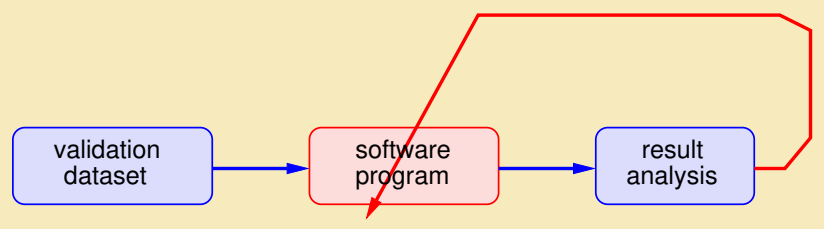
Nicola Angius and Alessio Plebe, 2023, **From Coding To Curing. Functions, Implementations, and Correctness in Deep Learning**



non-determinism from training

$$Y_j \leftarrow f_i(X_j)$$
$$Y_j \leftarrow f_i(X_j)$$
$$f_i \leftarrow \zeta(\Theta)$$

Nicola Angius and Alessio Plebe, 2023, **From Coding To Curing. Functions, Implementations, and Correctness in Deep Learning**





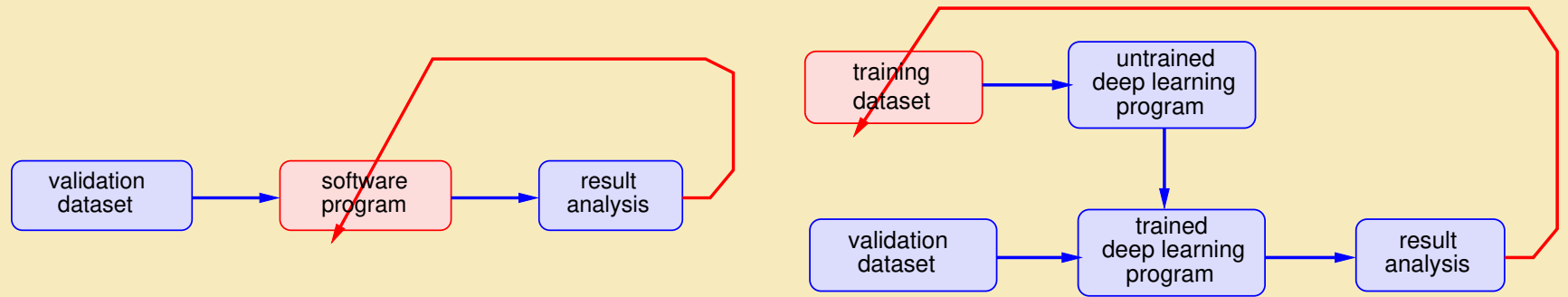
non-determinism from training

$$Y_j \leftarrow f_i(X_j)$$

$$Y_j \leftarrow f_i(X_j)$$

$$f_i \leftarrow \zeta(\Theta)$$

Nicola Angius and Alessio Plebe, 2023, **From Coding To Curing. Functions, Implementations, and Correctness in Deep Learning**





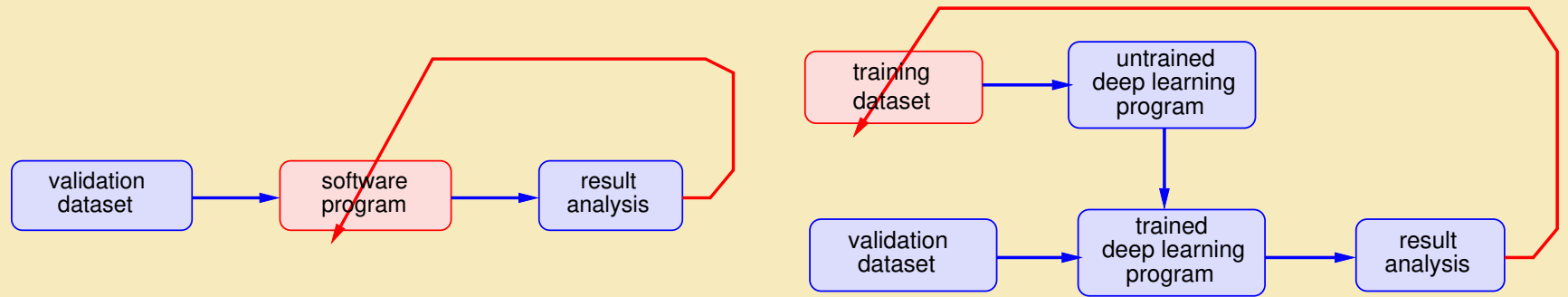
non-determinism from training

$$Y_j \leftarrow f_i(X_j)$$

$$Y_j \leftarrow f_i(X_j)$$

$$f_i \leftarrow \zeta(\Theta)$$

Nicola Angius and Alessio Plebe, 2023, **From Coding To Curing. Functions, Implementations, and Correctness in Deep Learning**





“universal” neurons

Wes Gurnee et al., 2024, **Universal neurons in GPT2 language models**

- ➔ 5 identical full training on GPT2 medium
- ➔ found 1253 universal neurons out the 98304
- ➔ distinct properties of universal neurons



“universal” neurons

Wes Gurnee et al., 2024, **Universal neurons in GPT2 language models**

- ➔ 5 identical full training on GPT2 medium
- ➔ found 1253 universal neurons out the 98304
- ➔ distinct properties of universal neurons



“universal” neurons

Wes Gurnee et al., 2024, **Universal neurons in GPT2 language models**

- ➔ 5 identical full training on GPT2 medium
- ➔ found 1253 universal neurons out the 98304
- ➔ distinct properties of universal neurons



“universal” neurons

Wes Gurnee et al., 2024, **Universal neurons in GPT2 language models**

- ➔ 5 identical full training on GPT2 medium
- ➔ found 1253 universal neurons out the 98304
- ➔ distinct properties of universal neurons



“universal” neurons

Wes Gurnee et al., 2024, **Universal neurons in GPT2 language models**

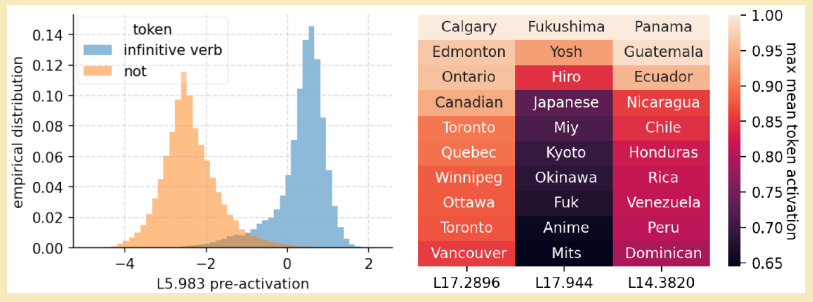
- ➔ 5 identical full training on GPT2 medium
- ➔ found 1253 universal neurons out the 98304
- ➔ distinct properties of universal neurons



“universal” neurons

Wes Gurnee et al., 2024, **Universal neurons in GPT2 language models**

- ➔ 5 identical full training on GPT2 medium
- ➔ found 1253 universal neurons out the 98304
- ➔ distinct properties of universal neurons

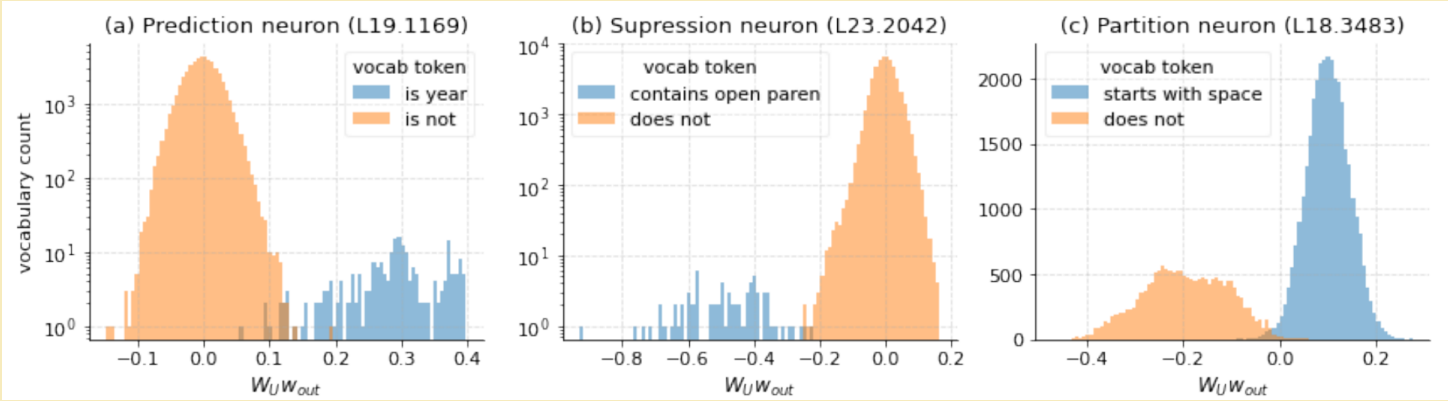
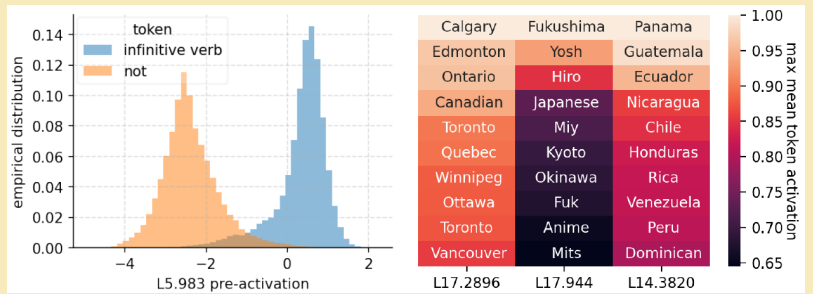




“universal” neurons

Wes Gurnee et al., 2024, **Universal neurons in GPT2 language models**

- ➔ 5 identical full training on GPT2 medium
- ➔ found 1253 universal neurons out the 98304
- ➔ distinct properties of universal neurons

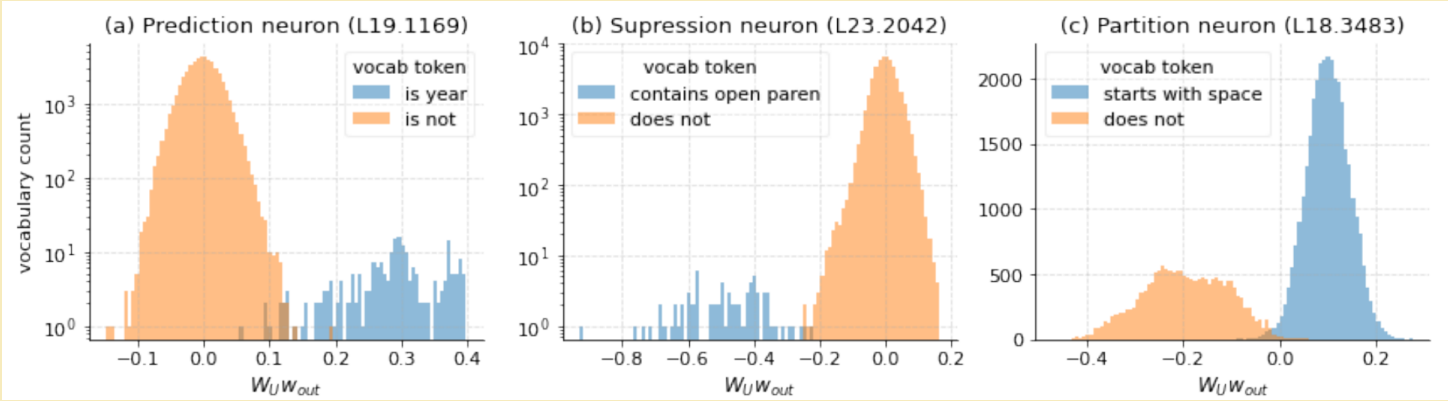
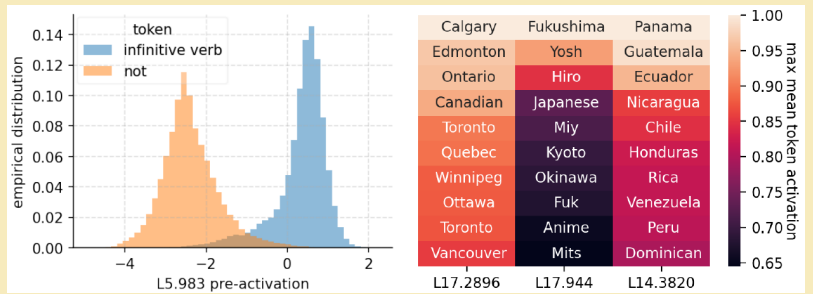




“universal” neurons

Wes Gurnee et al., 2024, **Universal neurons in GPT2 language models**

- ➔ 5 identical full training on GPT2 medium
- ➔ found 1253 universal neurons out the 98304
- ➔ distinct properties of universal neurons





thank you for your attention