

Language Models Need Inductive Biases to Count Inductively

Yingshan Chang and Yonatan Bisk

Why counting?

Why counting?

1. Counting is treated as a primitive in formal frameworks of Transformer's expressivity.

```
select_all = select(1,1,==)  
length = 1/aggregate(select_all, indicator(indices==0))
```

```
same_tok = select ( tokens, tokens, ==) ;  
hist = selector_width (  
    same_tok,  
    assume_bos = True  
)
```

Why counting?

1. Counting is treated as a primitive in formal frameworks of Transformer's expressivity.

```
select_all = select(1,1,==)  
length = 1/aggregate(select_all, indicator(indices==0))
```

```
same_tok = select ( tokens, tokens, ==) ;  
hist = selector_width (  
    same_tok,  
    assume_bos = True  
)
```

2. Counting is a key component in many complex tasks.

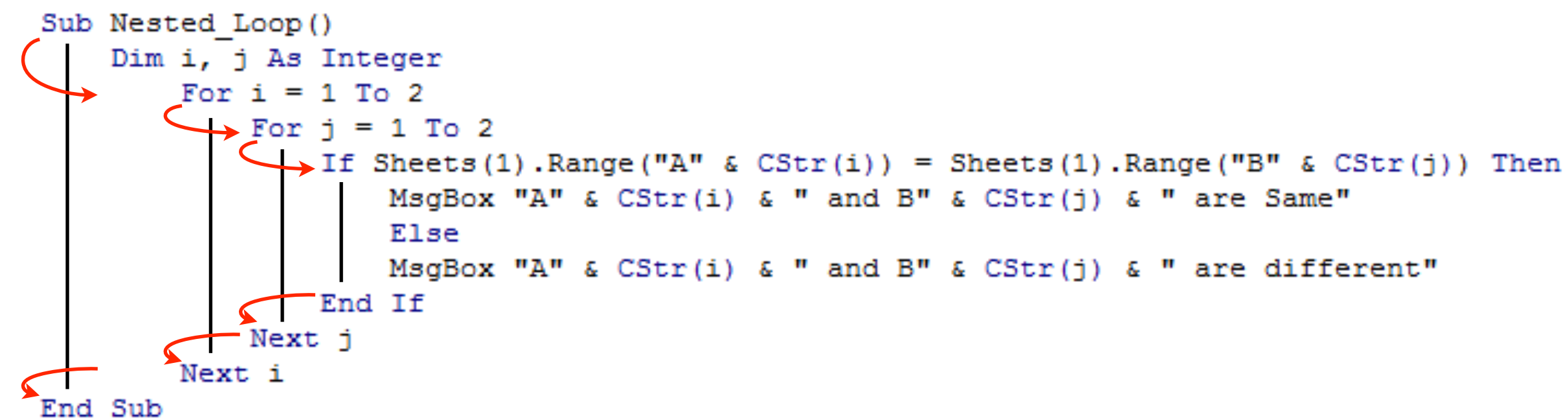
Why counting?

1. Counting is treated as a primitive in formal frameworks of Transformer's expressivity.

```
select_all = select(1,1,==)  
length = 1/aggregate(select_all, indicator(indices==0))
```

```
same_tok = select ( tokens, tokens, ==) ;  
hist = selector_width (  
    same_tok,  
    assume_bos = True  
)
```

2. Counting is a key component in many complex tasks.



```
Sub Nested_Loop()  
  Dim i, j As Integer  
  For i = 1 To 2  
    For j = 1 To 2  
      If Sheets(1).Range("A" & CStr(i)) = Sheets(1).Range("B" & CStr(j)) Then  
        MsgBox "A" & CStr(i) & " and B" & CStr(j) & " are Same"  
      Else  
        MsgBox "A" & CStr(i) & " and B" & CStr(j) & " are different"  
      End If  
    Next j  
  Next i  
End Sub
```

The diagram illustrates the execution flow of the 'Nested_Loop' subroutine. Red arrows indicate the sequence of operations: starting from the 'Sub' line, moving down to 'Dim', then to the 'For i' loop, then to the 'For j' loop, then to the 'If' statement, then to the 'MsgBox' statements, then to 'End If', then to 'Next j', then to 'Next i', and finally to 'End Sub'. The arrows show the nested nature of the loops and the conditional execution within the inner loop.

Why counting?

1. Counting is treated as a primitive in formal frameworks of Transformer's expressivity.

```
select_all = select(1,1,==)
length = 1/aggregate(select_all, indicator(indices==0))
```

```
same_tok = select ( tokens, tokens, ==) ;
hist = selector_width (
    same_tok,
    assume_bos = True
)
```

2. Counting is a key component in many complex tasks.

```
Sub Nested_Loop()
  Dim i, j As Integer
  For i = 1 To 2
    For j = 1 To 2
      If Sheets(1).Range("A" & CStr(j)) = "Bernard" Then
        MsgBox "A" & CStr(j)
      Else
        MsgBox "A" & CStr(j)
      End If
    Next j
  Next i
End Sub
```

Here is a table where the first line is a header and each subsequent line is a penguin:

name	age	height (cm)	weight (kg)
Louis	7	50	11
Bernard	5	80	13
Vincent	9	60	11
Gwen	8	70	15

For example: the age of Louis is 7, the weight of Gwen is 15 kg, the height of Bernard is 80 cm.

Which penguin is taller than the other ones? Answer:

Why counting?

1. Counting is treated as a primitive in formal frameworks of Transformer's expressivity.

```
select_all = select(1,1,==)
length = 1/aggregate(select_all, indicator(indices==0))
```

```
same_tok = select ( tokens, tokens, ==) ;
hist = selector_width (
    same_tok,
    assume_bos = True
)
```

2. Counting is a key component in many complex tasks.

```
Sub Nested_Loop()
  Dim i, j As Integer
  For i = 1 To 2
    For j = 1 To 2
      If Sheets(1).Range("A" & CStr(j)) = "A" Then
        MsgBox "A" & CStr(j)
      Else
        MsgBox "A" & CStr(j)
      End If
    Next j
  Next i
End Sub
```

Here is a table where the first line is a header and each

name	age	height (cm)	weight (kg)
Louis	7	50	11
Bernard	5	80	13
Vincent	9	60	11
Gwen	8	70	15

For example: the age of Louis is 7, the weight of Gwen is 15 kg, and the height of Bernard is 80 cm.

Which penguin is taller than the other ones? Answer:

Moreover, photosynthesis influences global climate patterns by removing carbon dioxide from the atmosphere, helping to regulate Earth's temperature. It also forms the basis of the food chain in ecosystems, providing energy for various organisms.

Understanding photosynthesis not only explains how life sustains itself through energy transfer, but also highlights the interdependence of living organisms and their environment. As such, efforts to mitigate climate change through sustainable agricultural practices or genetic engineering could have profound implications for food security and environmental health.

Furthermore, innovations in synthetic biology are pushing the boundaries of how we can harness biological processes. These advancements could lead to more efficient crops, capable of producing higher yields with fewer resources. This could revolutionize not only agriculture but energy sustainability, creating greener alternatives to fossil fuels.

300 words 2,031 characters

Why counting?

1. Counting is treated as a primitive in formal frameworks of Transformer's expressivity.

```
select_all = select(1,1,==)
length = 1/aggregate(select_all, indicator(indices==0))
```

```
same_tok = select ( tokens, tokens, ==) ;
hist = selector_width (
    same_tok,
    assume_bos = True
)
```

2. Counting is a key component in many complex tasks.

```
Sub Nested_Loop()
  Dim i, j As Integer
  For i = 1 To 2
    For j = 1 To 2
      If Sheets(1).Range("A" & CStr(i) & CStr(j)) = "A" Then
        MsgBox "A" & CStr(i) & CStr(j)
      Else
        MsgBox "A" & CStr(i) & CStr(j)
      End If
    Next j
  Next i
End Sub
```

Here is a table where the first line is a header and each

name	age	height (cm)	weight (kg)
Louis	7	50	11
Bernard	5	80	13
Vincent	9	60	11
Gwen	8	70	15

For example: the age of Louis is 7, the weight of Gwen is 15 kg, and the height of Bernard is 80 cm.

Which penguin is taller than the other ones? Answer:

Moreover, photosynthesis influences global climate patterns by removing carbon dioxide from the atmosphere, which helps to regulate Earth's temperature. It also forms the basis of the food chain in ecosystems, providing energy for various organisms.

Understanding photosynthesis not only explains how life sustains itself through energy transfer but also highlights the interdependence of living organisms and their environment. As such, efforts to improve agricultural practices or develop sustainable energy sources through genetic engineering could have profound implications for food security and environmental conservation.

Furthermore, innovations in synthetic biology are pushing the boundaries of how we can harness and mimic natural processes like photosynthesis. These advancements could lead to more efficient crops, capable of growing in harsher conditions or producing higher yields. This could revolutionize not only agriculture but energy sustainability, creating greener and more sustainable energy sources.

300 words 2,031 characters

3. The mechanism of counting is a *mapping* from set cardinalities to integers.

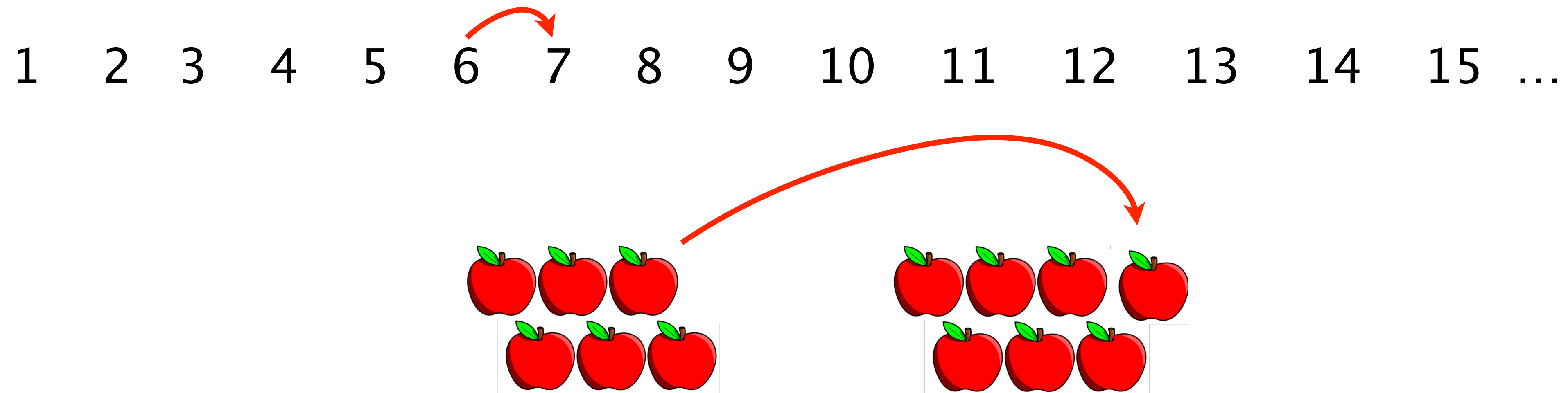
Why inductive counting?

Why inductive counting?

Counting inductively requires *extrapolating* the learned principle to larger counts.

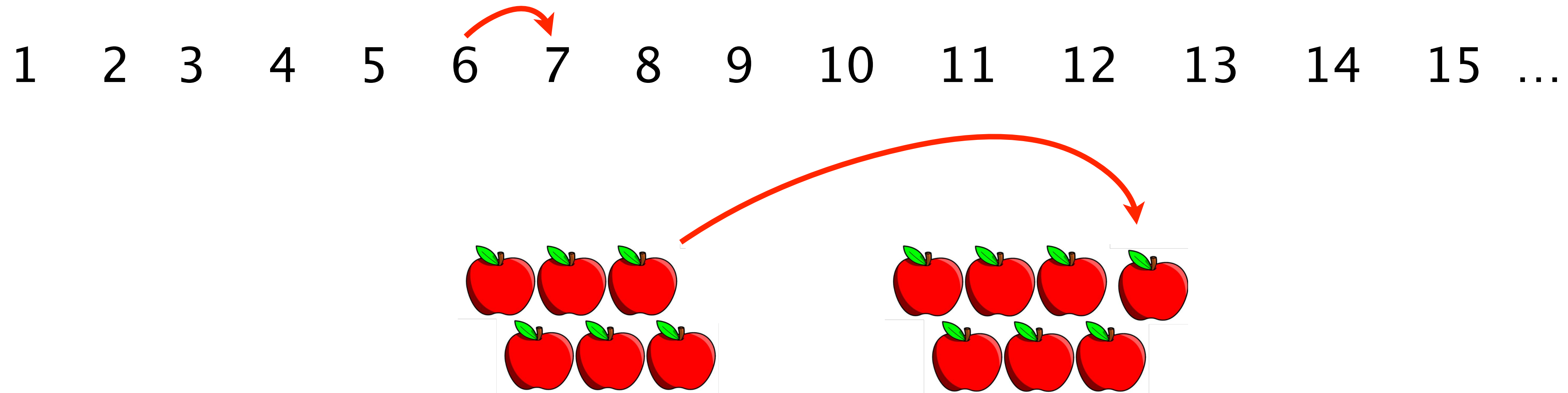
Why inductive counting?

Counting inductively requires *extrapolating* the learned principle to larger counts.



Why inductive counting?

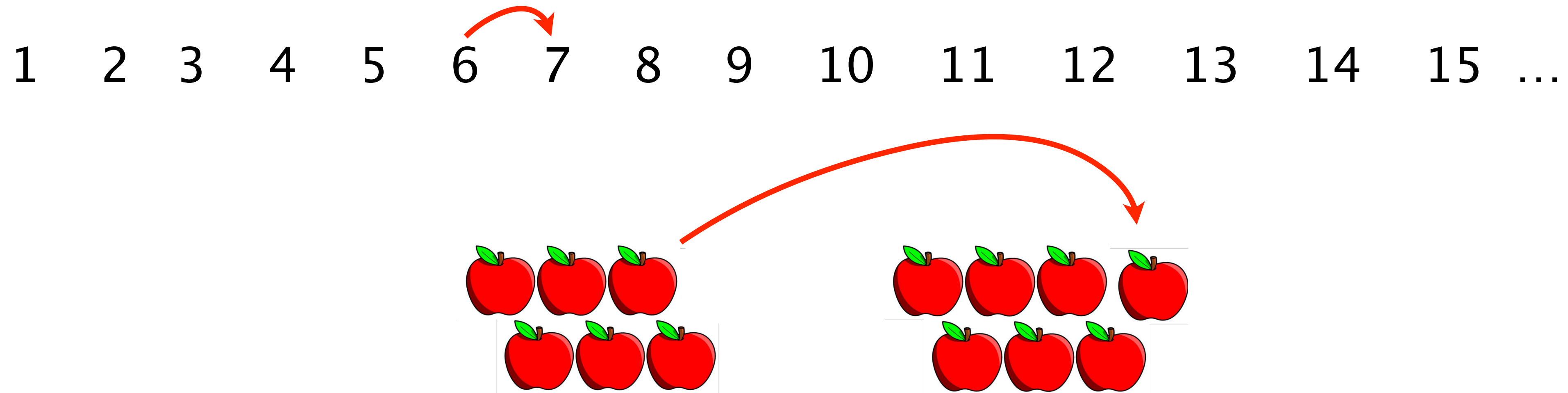
Counting inductively requires *extrapolating* the learned principle to larger counts.



The Inductive Counting Principle

Why inductive counting?

Counting inductively requires *extrapolating* the learned principle to larger counts.

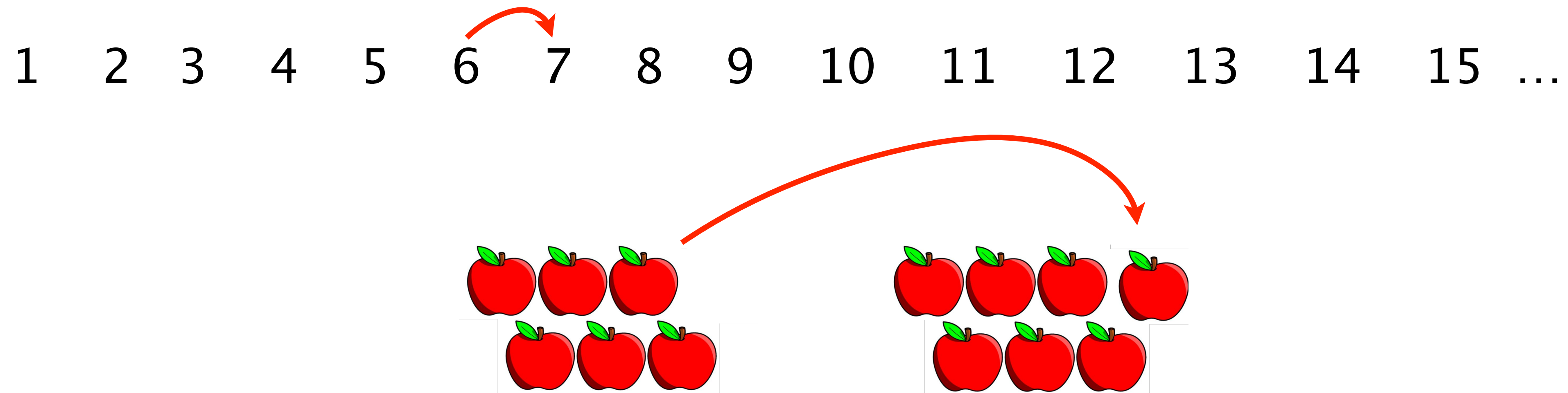


The Inductive Counting Principle

If a word in an ordered number word list refers to sets with cardinality n ,

Why *inductive* counting?

Counting inductively requires *extrapolating* the learned principle to larger counts.



The Inductive Counting Principle

If a word in an ordered number word list refers to sets with cardinality n ,
then the next word refers to sets with cardinality $n + 1$.

OOD cardinality is the hardest barrier

OOD cardinality is the hardest barrier

OOD Position 🤔

OOD cardinality is the hardest barrier

OOD Position 😞

OOD Vocabulary 😞

OOD cardinality is the hardest barrier

OOD Position 🙄

OOD Vocabulary 🙄

OOD Cardinality 🙄

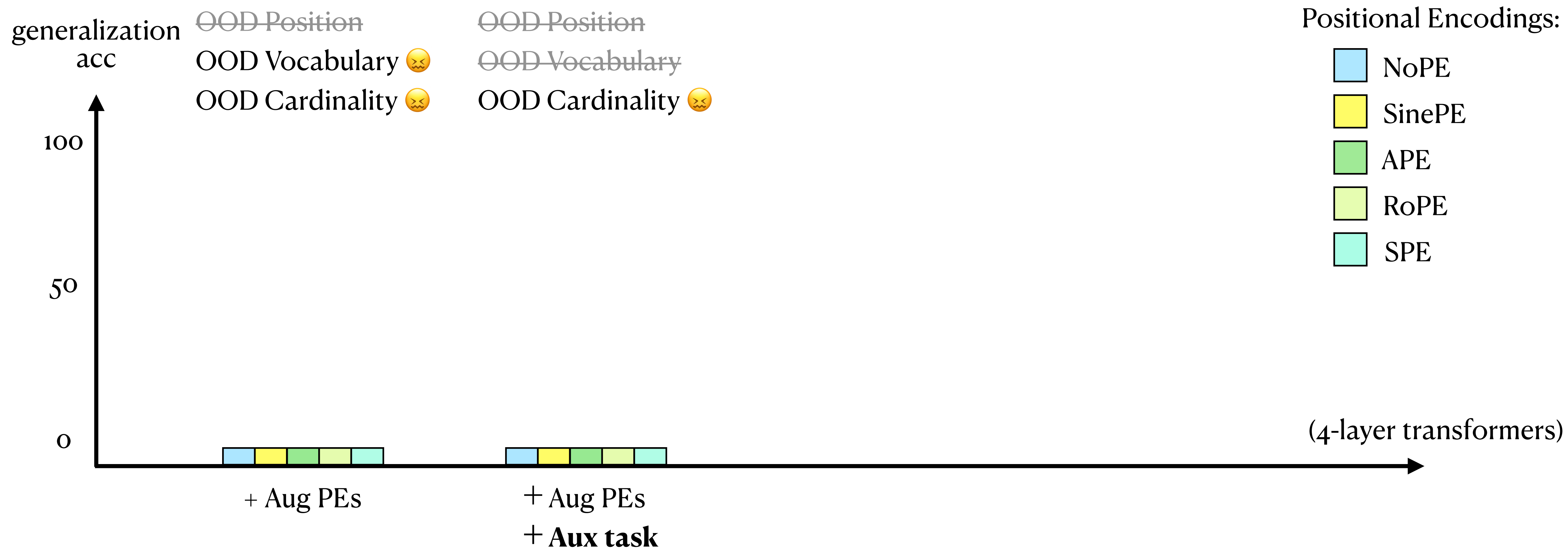
OOD cardinality is the hardest barrier



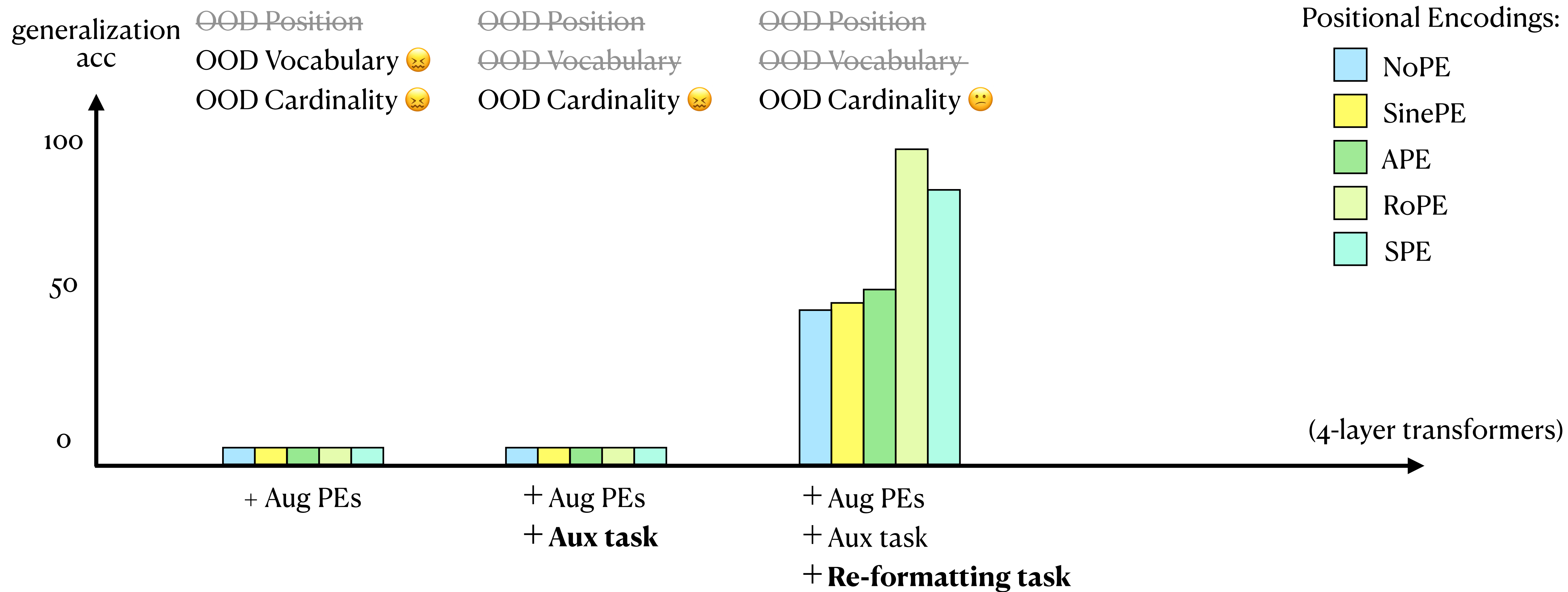
OOD cardinality is the hardest barrier



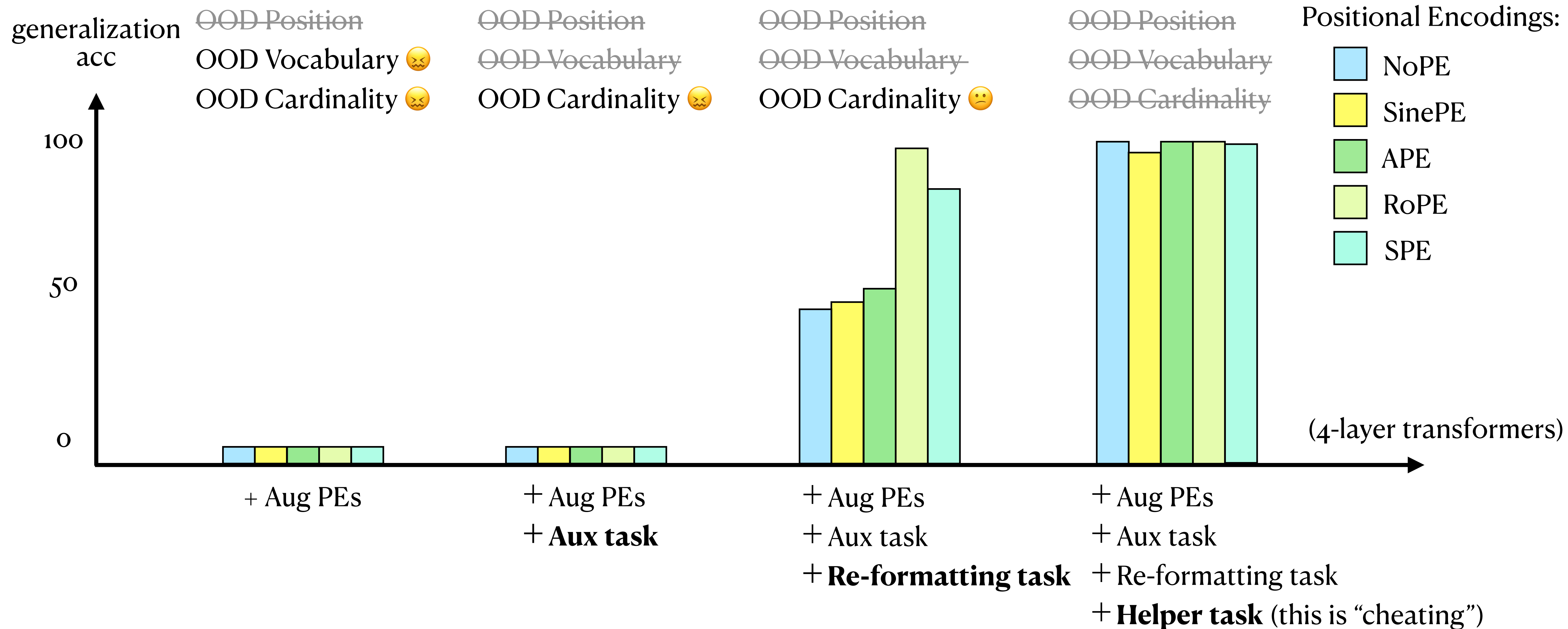
OOD cardinality is the hardest barrier



OOD cardinality is the hardest barrier




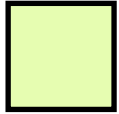
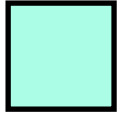


OOD cardinality is the hardest barrier



PEs provide the right inductive bias for modular and selective counting


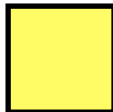

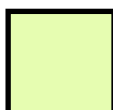
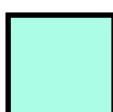
Positional Encodings:

-  NoPE
-  SinePE
-  APE
-  RoPE
-  SPE

PEs provide the right inductive bias for modular and selective counting

Modular counting

Positional Encodings:


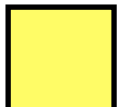


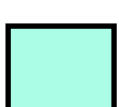
-  NoPE
-  SinePE
-  APE
-  RoPE
-  SPE

PEs provide the right inductive bias for modular and selective counting

Modular counting

First tok recog (homo)

Positional Encodings:

-  NoPE
-  SinePE
-  APE
-  RoPE
-  SPE


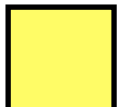


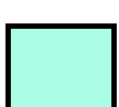
PEs provide the right inductive bias for modular and selective counting

Modular counting

First tok recog (homo)

Circular PE structure

Positional Encodings:

-  NoPE
-  SinePE
-  APE
-  RoPE
-  SPE

PEs provide the right inductive bias for modular and selective counting


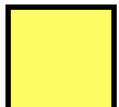


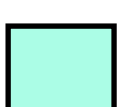
Modular counting

First tok recog (homo)

Circular PE structure

Selective counting

Positional Encodings:

-  NoPE
-  SinePE
-  APE
-  RoPE
-  SPE

PEs provide the right inductive bias for modular and selective counting

Modular counting


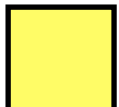


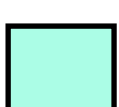
First tok recog (homo)

Circular PE structure

Selective counting

First tok recog (hetero)

Positional Encodings:

-  NoPE
-  SinePE
-  APE
-  RoPE
-  SPE

PEs provide the right inductive bias for modular and selective counting

Modular counting

First tok recog (homo)



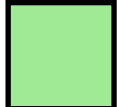


Circular PE structure

Selective counting

First tok recog (hetero)

Content-based attn

Positional Encodings:

-  NoPE
-  SinePE
-  APE
-  RoPE
-  SPE















PEs provide the right inductive bias for modular and selective counting

Modular counting		<div><div></div><div></div><div></div><div></div><div></div></div>					Selective counting	
First tok recog (homo)	×	✓	✓	×	✓		First tok recog (hetero)	
Circular PE structure	×	✓	✓	×	×		Content-based attn	

Positional Encodings:

- NoPE
- SinePE
- APE
- RoPE
- SPE

PEs provide the right inductive bias for modular and selective counting

Modular counting						Selective counting								
													SinePE	
First tok recog (homo)	×	✓	✓	×	✓		First tok recog (hetero)	✓	✓	✓	×	✓		APE
Circular PE structure	×	✓	✓	×	×		Content-based attn	✓	✓	✓	×	✓		RoPE
													SPE	

PEs provide the right inductive bias for modular and selective counting

Modular counting

First tok recog (homo)

Circular PE structure



×	✓	✓	×	✓
×	✓	✓	×	×

Selective counting

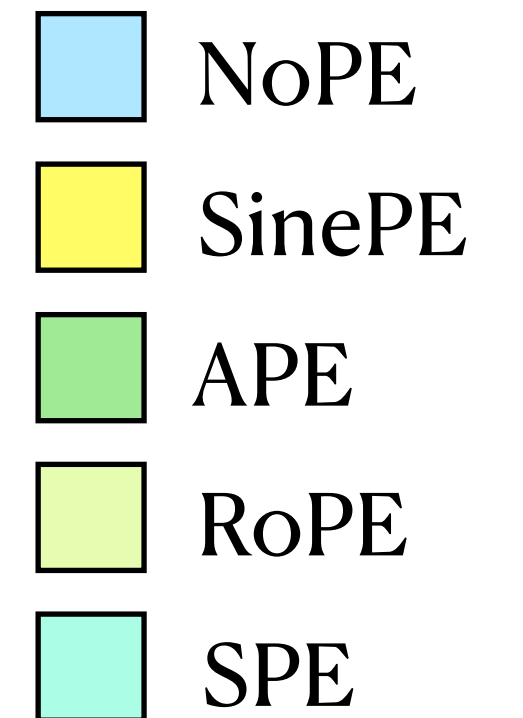
First tok recog (hetero)

Content-based attn



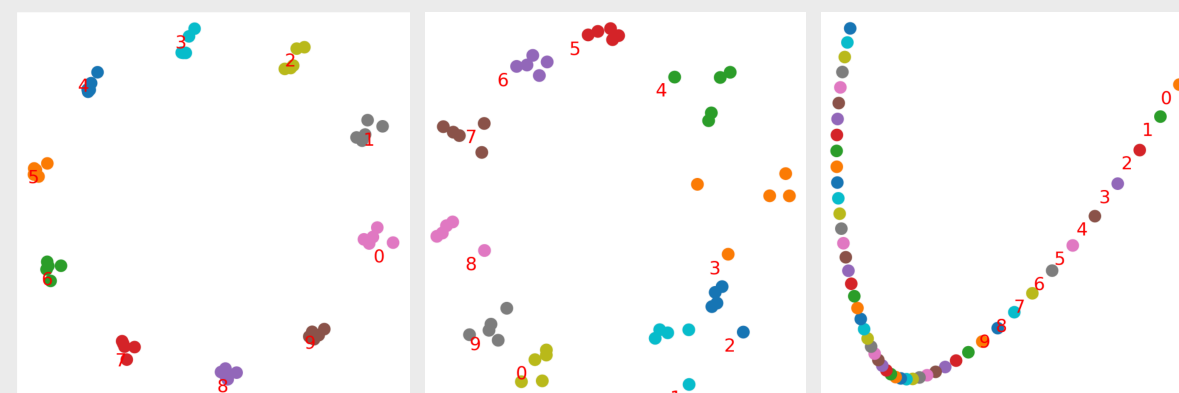
✓	✓	✓	×	✓
✓	✓	✓	×	✓

Positional Encodings:

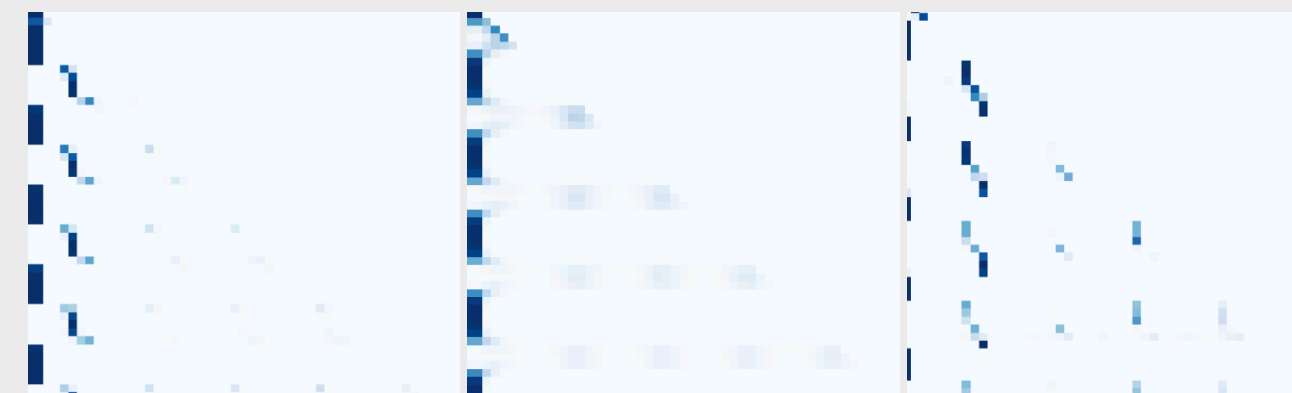


Mechanistic Analyses

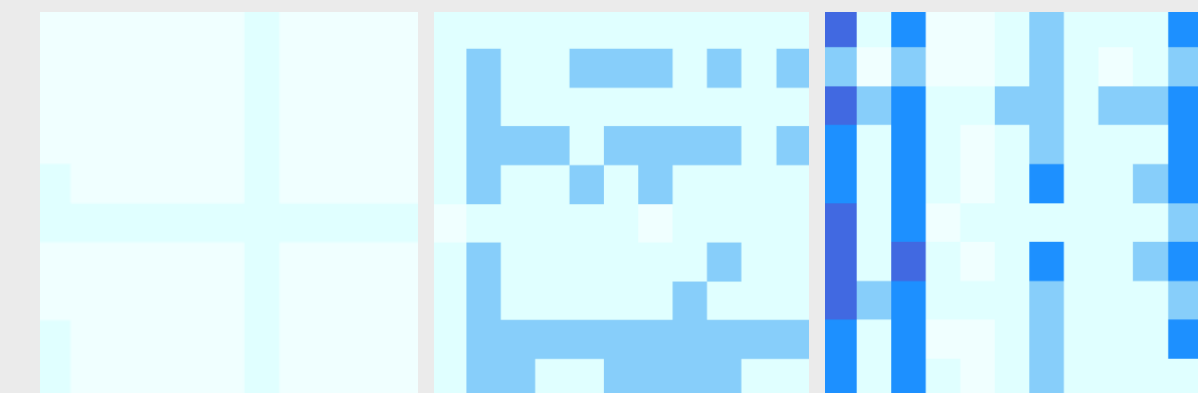
PCA of internal representations



Attention weights

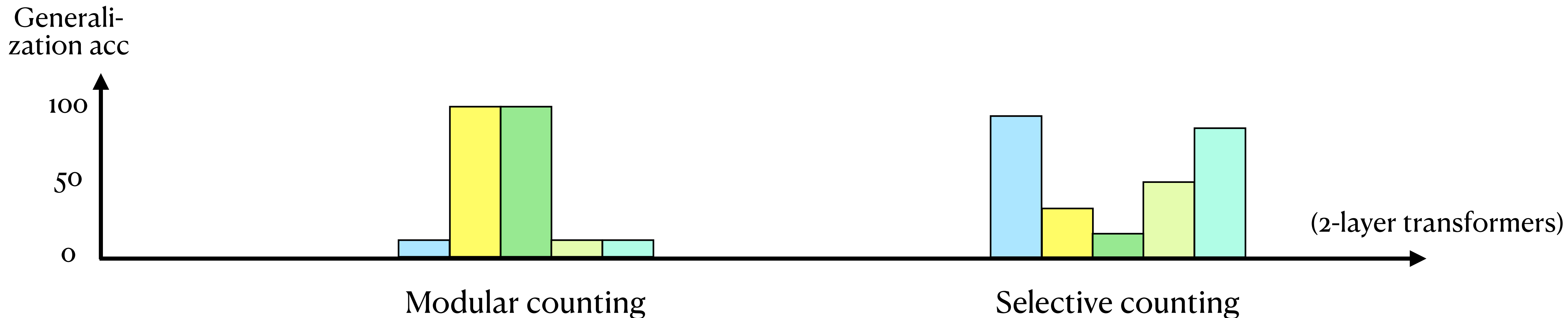


Attention strength variation
over same-identity tokens



PEs provide the right inductive bias for modular and selective counting

	Positional Encodings:						Positional Encodings:				
	NoPE	SinePE	APE	RoPE	SPE		NoPE	SinePE	APE	RoPE	SPE
Modular counting						Selective counting					
First tok recog (homo)	×	✓	✓	×	✓	First tok recog (hetero)	✓	✓	✓	×	✓
Circular PE structure	×	✓	✓	×	×	Content-based attn	✓	✓	✓	×	✓



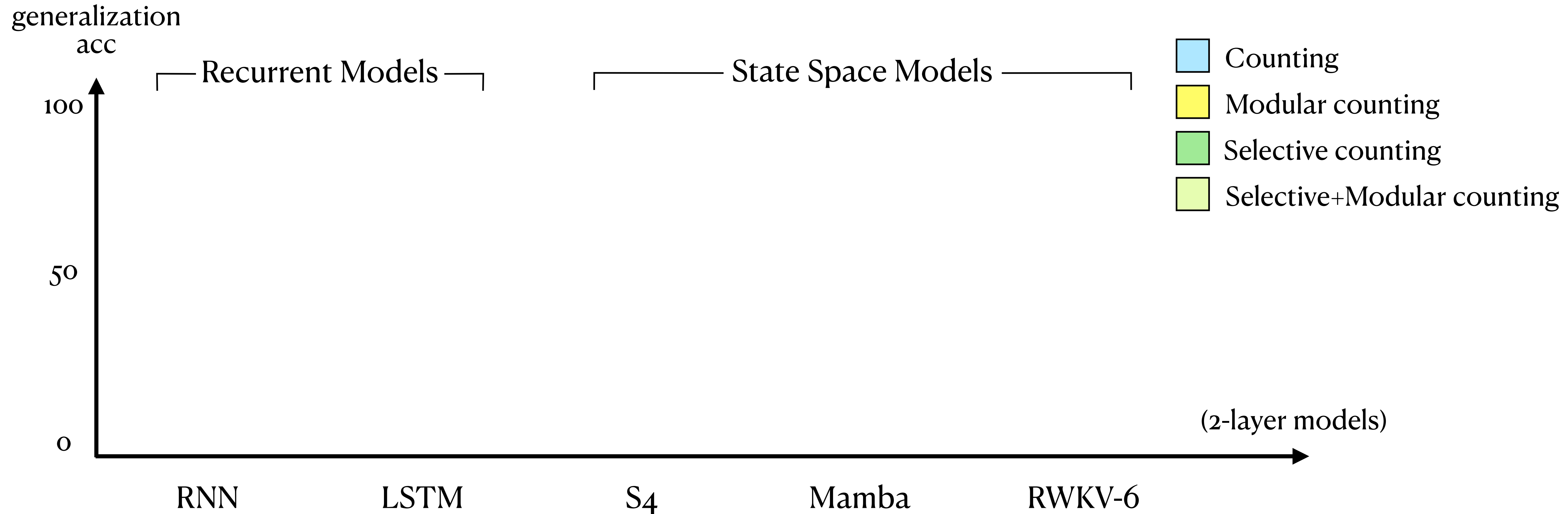
What about other architectures?



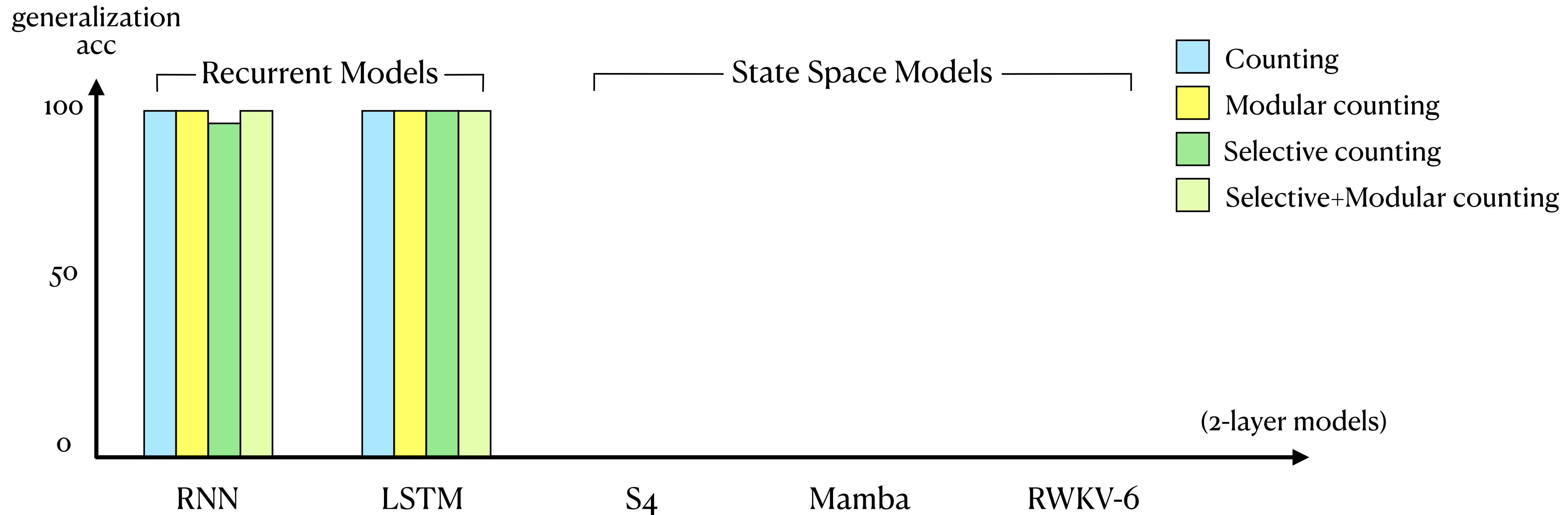
What about other architectures?



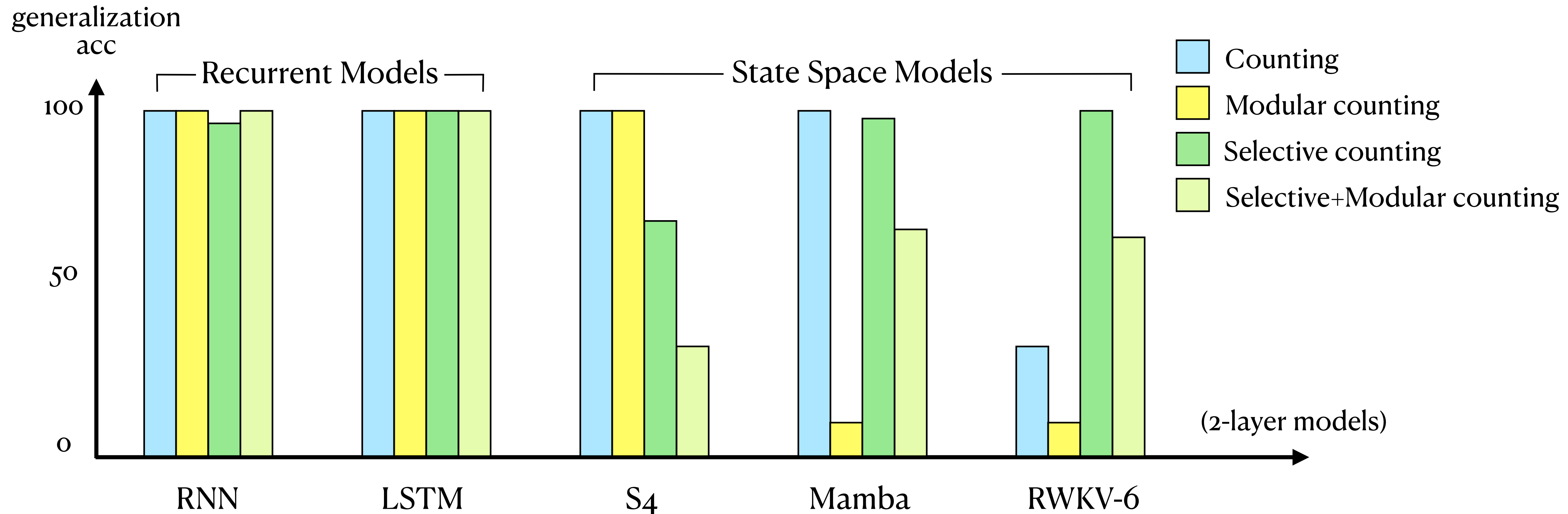
What about other architectures?



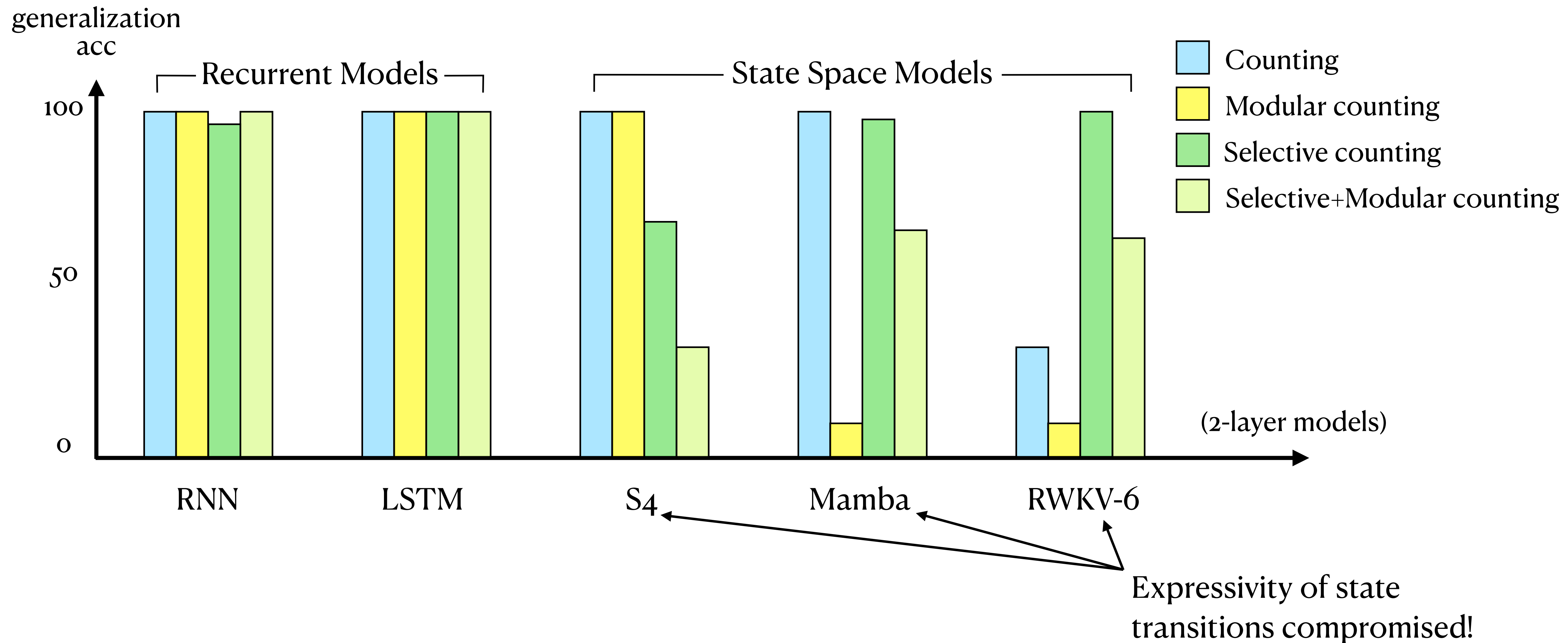
What about other architectures?



What about other architectures?



What about other architectures?



Conclusion

1. The difficulty of inductive counting:
 - not about tackling unseen vocabularies
 - not about tackling unseen positions
 - It is about tackling unseen **cardinalities**

Conclusion

1. The difficulty of inductive counting:
 - not about tackling unseen vocabularies
 - not about tackling unseen positions
 - It is about tackling unseen **cardinalities**
2. Architecture choice should inform desired **inductive bias**.

Conclusion

1. The difficulty of inductive counting:
 - not about tackling unseen vocabularies
 - not about tackling unseen positions
 - It is about tackling unseen **cardinalities**
2. Architecture choice should inform desired **inductive bias**.
3. Causes of length-generalization failures are **multi-faceted**.