Initialization Is Critical to Reasoning Ability of Transformer

Zhongwang Zhang
Shanghai Jiao Tong University

Wednesday, November 13, 2024

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Inference example

Try to remember these equations and do test, e.g.,（2，1，24）=?

$$(1,2,21)=27,(1,2,30)=36,(1,2,47)=53,(2,1,15)=21,(2,1,24)=30,(2,1,41)=47$$

$$(3,4,10)=0,(3,4,19)=9,(3,4,34)=24,(3,4,41)=31,(4,3,18)=8,(4,3,27)=17$$

$$(1,3,14)=17,(1,3,18)=21,(1,3,27)=30,(1,3,35)=38,(3,1,9)=12,(3,1,13)=16$$

$$(2,4,16)=9,(2,4,17)=10,(2,4,30)=23,(2,4,38)=31,(4,2,13)=6,(4,2,14)=7$$

$$(1,4,32)=29,(1,4,26)=23,(1,4,13)=10,(1,4,6)=3,(4,1,32)=29,(4,1,26)=23$$

$$(2,3,22)=21,(2,3,17)=16,(2,3,28)=27,(2,3,12)=11,(3,2,22)=21,(3,2,17)=16$$

$$\vdots$$

Inference example

$$(1,2,x) = x+6$$
$$(2,1,x) = x+6$$
$$(1,3,x) = x+3$$
$$(3,1,x) = x+3$$
$$(1,4,x) = x-3$$
$$(4,1,x) = x-3$$
$$(3,2,x) = x-1$$
$$(2,2,x) = x-1$$
$$(4,2,x) = x-9$$
$$(2,4,x) = x-9$$
$$(3,4,x) = x+6$$
$$(4,3,x) = x+6$$
$$(1,1,x) = x+10$$
$$(2,2,x) = x+2$$
$$(3,3,x) = x-4$$
$$(4,4,x) = x-16$$

$$(1,2,21) = 27, \quad (2,1,29) = 35$$
$$(2,1,50) = 56, \quad (1,2,31) = 37$$
$$(1,3,16) = 19, \quad (3,1,16) = 19$$
$$(1,4,67) = 64, \quad (4,1,99) = 96$$
$$(3,2,60) = 59, \quad (2,3,50) = 49$$
$$(4,2,48) = 39, \quad (2,4,33) = 24$$
$$(3,4,77) = 83, \quad (4,3,90) = 96$$
$$(1,1,58) = 68, \quad (1,1,51) = 61$$
$$(2,2,46) = 48, \quad (2,2,35) = 37$$
$$(3,3,36) = 32, \quad (3,3,29) = 25$$
$$(4,4,88) = 54, \quad (4,4,46) = 30$$

$$\vdots$$

Inference example

$$
\begin{aligned}
(1,2,x) &= (2,1,x) = x+6 \\
(1,3,x) &= (3,1,x) = x+3 \\
(1,4,x) &= (4,1,x) = x-3 \\
(3,2,x) &= (2,2,x) = x-1 \\
(4,2,x) &= (2,4,x) = x-9 \\
(3,4,x) &= (4,3,x) = x+6 \\
(1,1,x) &= x+10 \\
(2,2,x) &= x+2 \\
(3,3,x) &= x-4 \\
(4,4,x) &= x-16
\end{aligned}
$$

$$
\begin{aligned}
(1,2,x) &= x+6 \\
(2,1,x) &= x+6 \\
(1,3,x) &= x+3 \\
(3,1,x) &= x+3 \\
(1,4,x) &= x-3 \\
(4,1,x) &= x-3 \\
(3,2,x) &= x-1 \\
(2,2,x) &= x-1 \\
(4,2,x) &= x-9 \\
(2,4,x) &= x-9 \\
(3,4,x) &= x+6 \\
(4,3,x) &= x+6 \\
(1,1,x) &= x+10 \\
(2,2,x) &= x+2 \\
(3,3,x) &= x-4 \\
(4,4,x) &= x-16
\end{aligned}
$$

$$
\begin{aligned}
(1,2,21) &= 27, & (2,1,29) &= 35 \\
(2,1,50) &= 56, & (1,2,31) &= 37 \\
(1,3,16) &= 19, & (3,1,16) &= 19 \\
(1,4,67) &= 64, & (4,1,99) &= 96 \\
(3,2,60) &= 59, & (2,3,50) &= 49 \\
(4,2,48) &= 39, & (2,4,33) &= 24 \\
(3,4,77) &= 83, & (4,3,90) &= 96 \\
(1,1,58) &= 68, & (1,1,51) &= 61 \\
(2,2,46) &= 48, & (2,2,35) &= 37 \\
(3,3,36) &= 32, & (3,3,29) &= 25 \\
(4,4,88) &= 54, & (4,4,46) &= 30
\end{aligned}
$$

$$\vdots$$

Inference example

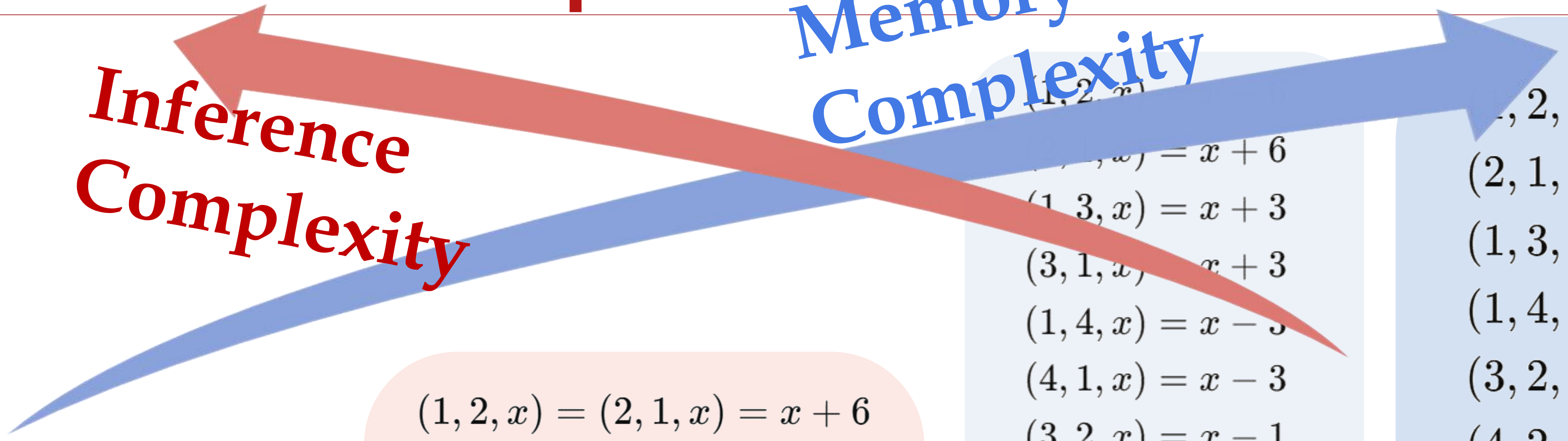
$$f_1(x) = x + 5$$
$$f_2(x) = x + 1$$
$$f_3(x) = x - 2$$
$$f_4(x) = x - 8$$

$$(1,2,x) = (2,1,x) = x + 6$$
$$(1,3,x) = (3,1,x) = x + 3$$
$$(1,4,x) = (4,1,x) = x - 3$$
$$(3,2,x) = (2,2,x) = x - 1$$
$$(4,2,x) = (2,4,x) = x - 9$$
$$(3,4,x) = (4,3,x) = x + 6$$
$$(1,1,x) = x + 10$$
$$(2,2,x) = x + 2$$
$$(3,3,x) = x - 4$$
$$(4,4,x) = x - 16$$

$$(1,2,x) = x + 6$$
$$(2,1,x) = x + 6$$
$$(1,3,x) = x + 3$$
$$(3,1,x) = x + 3$$
$$(1,4,x) = x - 3$$
$$(4,1,x) = x - 3$$
$$(3,2,x) = x - 1$$
$$(2,2,x) = x - 1$$
$$(4,2,x) = x - 9$$
$$(2,4,x) = x - 9$$
$$(3,4,x) = x + 6$$
$$(4,3,x) = x + 6$$
$$(1,1,x) = x + 10$$
$$(2,2,x) = x + 2$$
$$(3,3,x) = x - 4$$
$$(4,4,x) = x - 16$$

$$(1,2,21) = 27, \quad (2,1,29) = 35$$
$$(2,1,50) = 56, \quad (1,2,31) = 37$$
$$(1,3,16) = 19, \quad (3,1,16) = 19$$
$$(1,4,67) = 64, \quad (4,1,99) = 96$$
$$(3,2,60) = 59, \quad (2,3,50) = 49$$
$$(4,2,48) = 39, \quad (2,4,33) = 24$$
$$(3,4,77) = 83, \quad (4,3,90) = 96$$
$$(1,1,58) = 68, \quad (1,1,51) = 61$$
$$(2,2,46) = 48, \quad (2,2,35) = 37$$
$$(3,3,36) = 32, \quad (3,3,29) = 25$$
$$(4,4,88) = 54, \quad (4,4,46) = 30$$
$$\vdots$$

Inference example



$$f_1(x) = x + 5$$
$$f_2(x) = x + 1$$
$$f_3(x) = x - 2$$
$$f_4(x) = x - 8$$

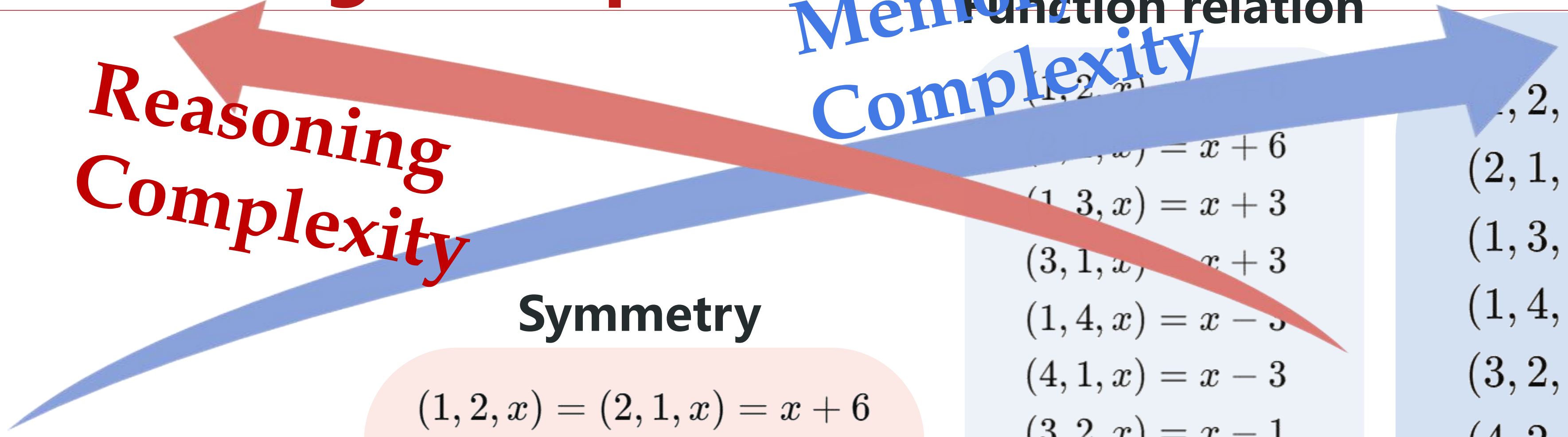
$$(1,2,x) = (2,1,x) = x + 6$$
$$(1,3,x) = (3,1,x) = x + 3$$
$$(1,4,x) = (4,1,x) = x - 3$$
$$(3,2,x) = (2,2,x) = x - 1$$
$$(4,2,x) = (2,4,x) = x - 9$$
$$(3,4,x) = (4,3,x) = x + 6$$
$$(1,1,x) = x + 10$$
$$(2,2,x) = x + 2$$
$$(3,3,x) = x - 4$$
$$(4,4,x) = x - 16$$

$$(1,2,x)$$
$$(2,1,x) = x + 6$$
$$(1,3,x) = x + 3$$
$$(3,1,x) = x + 3$$
$$(1,4,x) = x - 3$$
$$(4,1,x) = x - 3$$
$$(3,2,x) = x - 1$$
$$(2,2,x) = x - 1$$
$$(4,2,x) = x - 9$$
$$(2,4,x) = x - 9$$
$$(3,4,x) = x + 6$$
$$(4,3,x) = x + 6$$
$$(1,1,x) = x + 10$$
$$(2,2,x) = x + 2$$
$$(3,3,x) = x - 4$$
$$(4,4,x) = x - 16$$

$$(1,2,21) = 27, \quad (2,1,29) = 35$$
$$(2,1,50) = 56, \quad (1,2,31) = 37$$
$$(1,3,16) = 19, \quad (3,1,16) = 19$$
$$(1,4,67) = 64, \quad (4,1,99) = 96$$
$$(3,2,60) = 59, \quad (2,3,50) = 49$$
$$(4,2,48) = 39, \quad (2,4,33) = 24$$
$$(3,4,77) = 83, \quad (4,3,90) = 96$$
$$(1,1,58) = 68, \quad (1,1,51) = 61$$
$$(2,2,46) = 48, \quad (2,2,35) = 37$$
$$(3,3,36) = 32, \quad (3,3,29) = 25$$
$$(4,4,88) = 54, \quad (4,4,46) = 30$$
$$\vdots$$

Reasoning example

How will LLM do?



Composition

$f_1(x) = x + 5$

$f_2(x) = x + 1$

$f_3(x) = x - 2$

$f_4(x) = x - 8$

Symmetry

$(1, 2, x) = (2, 1, x) = x + 6$

$(1, 3, x) = (3, 1, x) = x + 3$

$(1, 4, x) = (4, 1, x) = x - 3$

$(3, 2, x) = (2, 2, x) = x - 1$

$(4, 2, x) = (2, 4, x) = x - 9$

$(3, 4, x) = (4, 3, x) = x + 6$

$(1, 1, x) = x + 10$

$(2, 2, x) = x + 2$

$(3, 3, x) = x - 4$

$(4, 4, x) = x - 16$

Function relation

$(1, 2, x)$

$) = x + 6$

$(1, 3, x) = x + 3$

$(3, 1, x) = x + 3$

$(1, 4, x) = x - 3$

$(4, 1, x) = x - 3$

$(3, 2, x) = x - 1$

$(2, 2, x) = x - 1$

$(4, 2, x) = x - 9$

$(2, 4, x) = x - 9$

$(3, 4, x) = x + 6$

$(4, 3, x) = x + 6$

$(1, 1, x) = x + 10$

$(2, 2, x) = x + 2$

$(3, 3, x) = x - 4$

$(4, 4, x) = x - 16$

Memorizing

$, 2, 21) = 27, \quad (2, 1, 29) = 35$

$(2, 1, 50) = 56, \quad (1, 2, 31) = 37$

$(1, 3, 16) = 19, \quad (3, 1, 16) = 19$

$(1, 4, 67) = 64, \quad (4, 1, 99) = 96$

$(3, 2, 60) = 59, \quad (2, 3, 50) = 49$

$(4, 2, 48) = 39, \quad (2, 4, 33) = 24$

$(3, 4, 77) = 83, \quad (4, 3, 90) = 96$

$(1, 1, 58) = 68, \quad (1, 1, 51) = 61$

$(2, 2, 46) = 48, \quad (2, 2, 35) = 37$

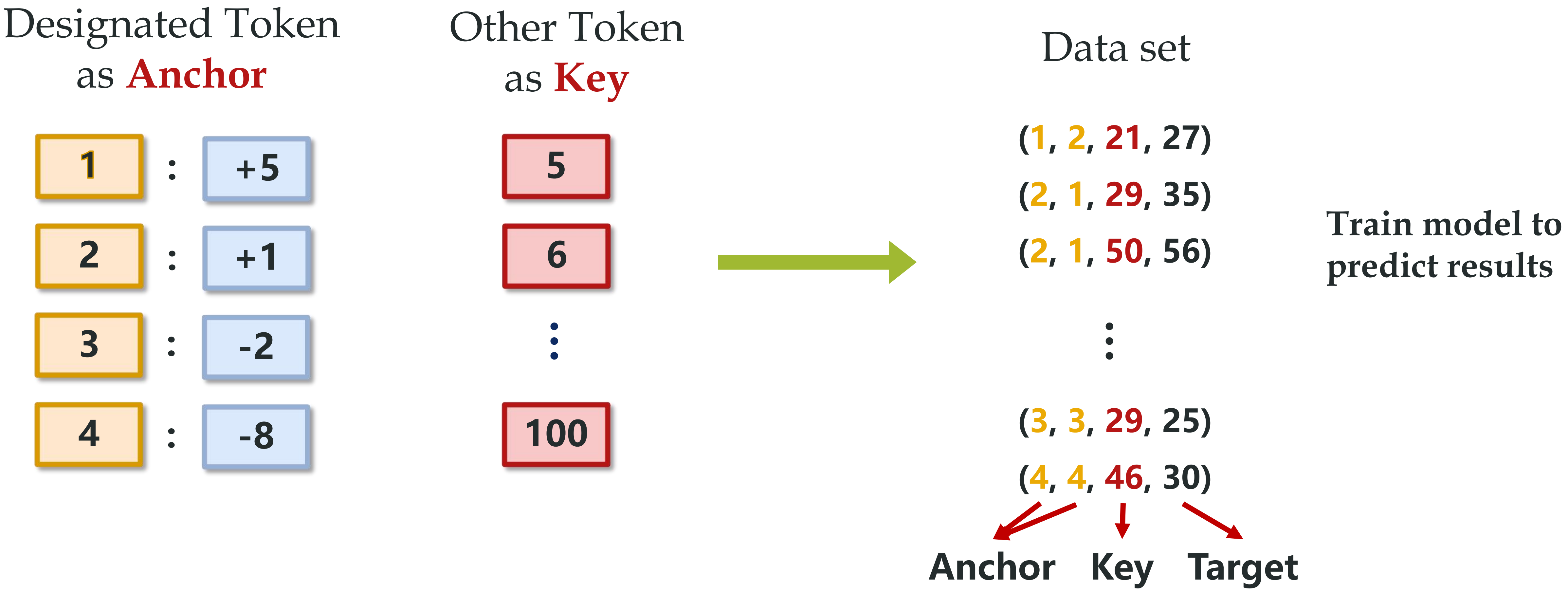
$(3, 3, 36) = 32, \quad (3, 3, 29) = 25$

$(4, 4, 88) = 54, \quad (4, 4, 46) = 30$

$\vdots$

Zhongwang Zhang, Pengxiao Lin, Zhiwei Wang, Yaoyu Zhang, Zhi-Qin John Xu*, Initialization is Critical to Whether Transformers Fit Composite Functions by Inference or Memorizing, NeurIPS 2024

Anchor function



Designated Token as **Anchor**

1 : +5

2 : +1

3 : -2

4 : -8

Other Token as **Key**

5

6

⋮

100

Data set

(1, 2, 21, 27)

(2, 1, 29, 35)

(2, 1, 50, 56)

⋮

(3, 3, 29, 25)

(4, 4, 46, 30)

Anchor Key Target

Train model to predict results

NN to learn Anchor function

Data set

(1, 2, 21, 27)

(2, 1, 29, 35)

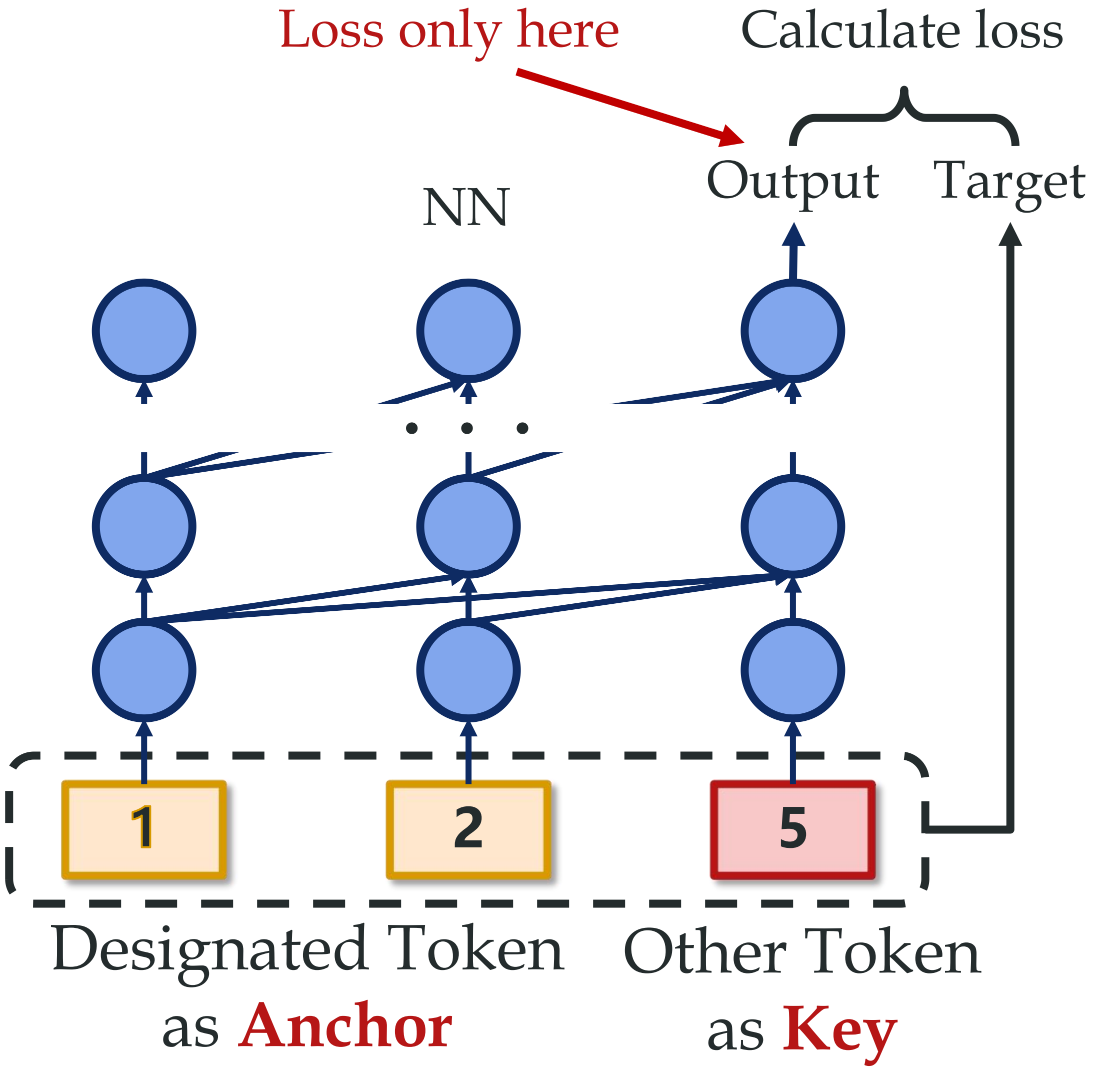
(2, 1, 50, 56)

⋮

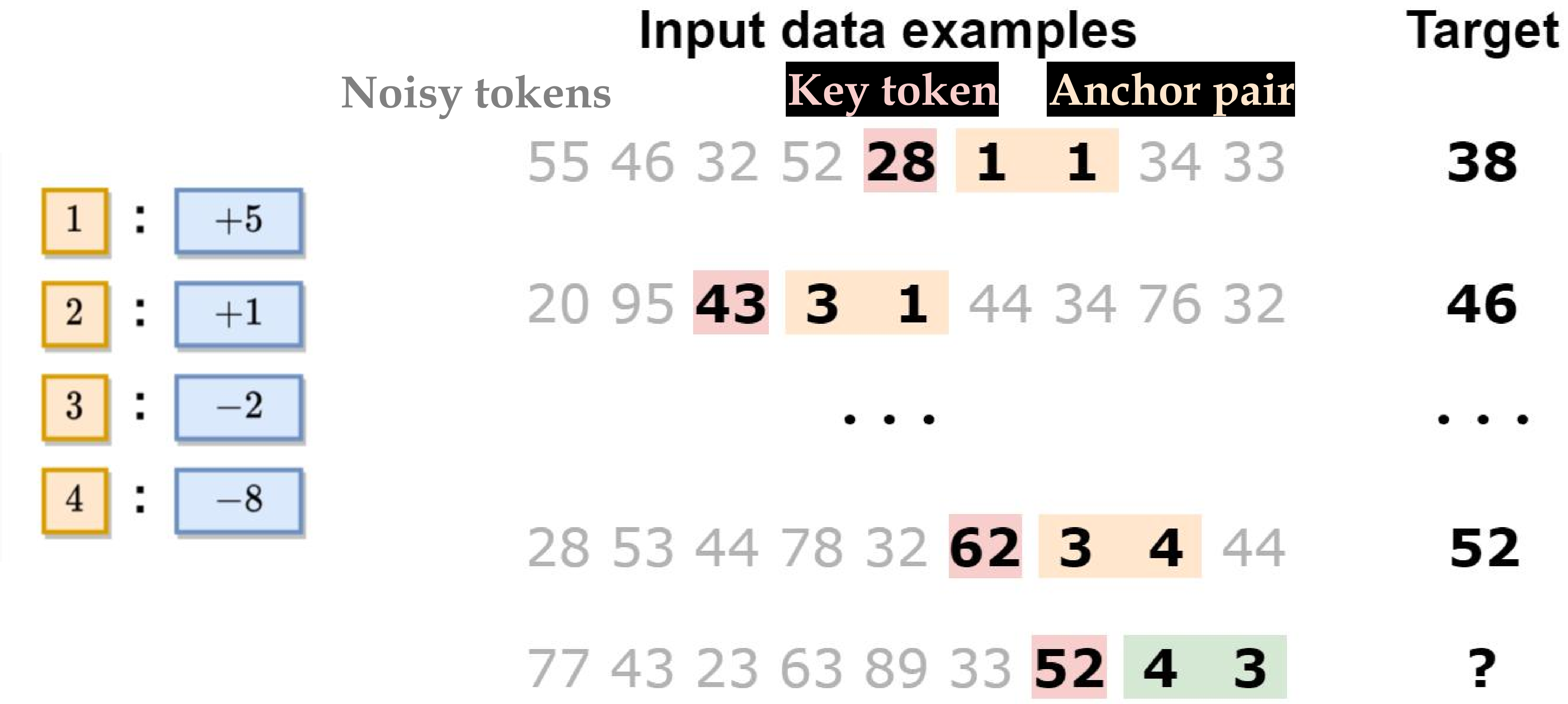
(3, 3, 29, 25)

(4, 4, 46, 30)

Anchor Key Target



Composite Anchor function



1	:	+5
2	:	+1
3	:	−2
4	:	−8

Input data examples	Target
Noisy tokens / Key token / Anchor pair	
55 46 32 52 **28** **1 1** 34 33	**38**
20 95 **43** **3 1** 44 34 76 32	**46**
. . .	. . .
28 53 44 78 32 **62** **3 4** 44	**52**
77 43 23 63 89 33 **52** **4 3**	**?**

Zhongwang Zhang, Pengxiao Lin, Zhiwei Wang, Yaoyu Zhang, Zhi-Qin John Xu*, Initialization is Critical to Whether Transformers Fit Composite Functions by Inference or Memorizing, NeurIPS 2024

Can it learn [4,3]?

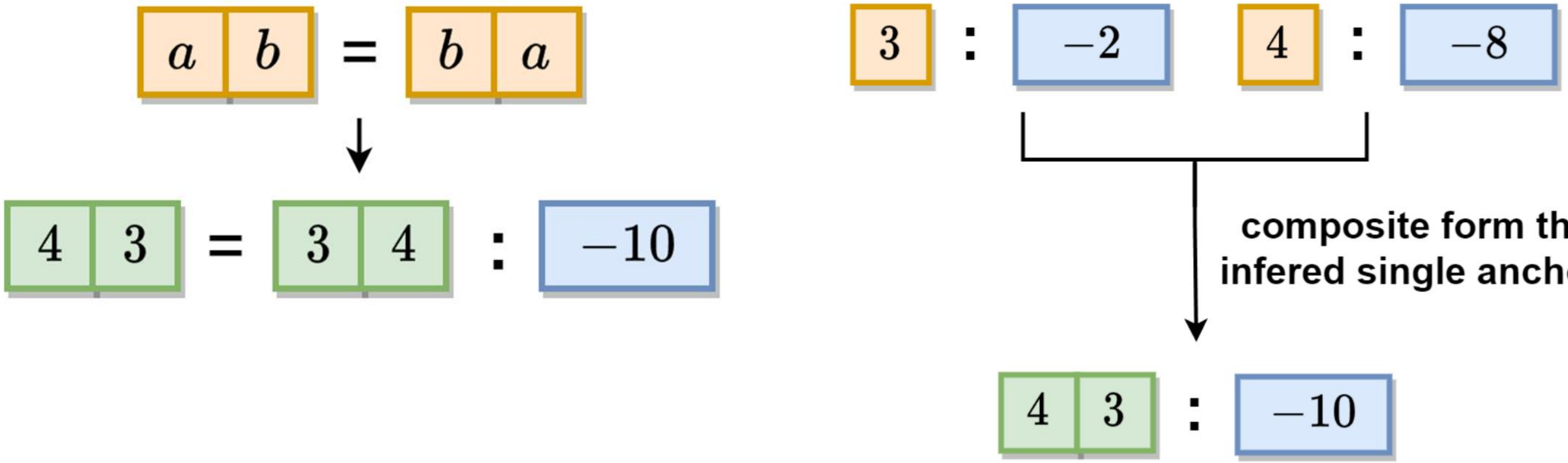
Yes, it can learn [4,3]!

Turn your eyes to preferred answer

Symmetric or Inferential?

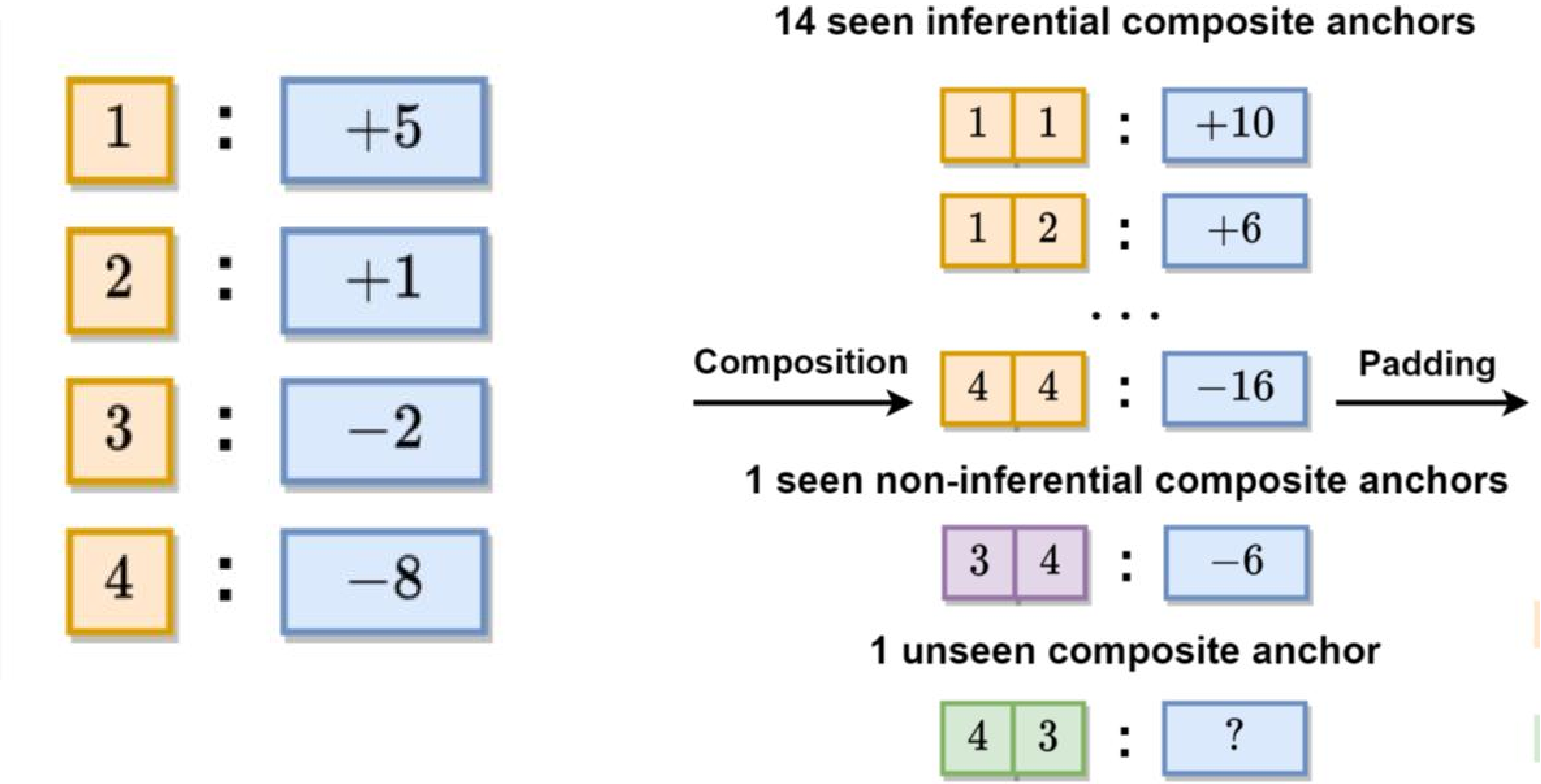
Mechanism 1: learn symmetric structure

Mechanism 2: infer single anchor mappings



composite form the infered single anchors

Composite Anchor function



14 seen inferential composite anchors

Composition

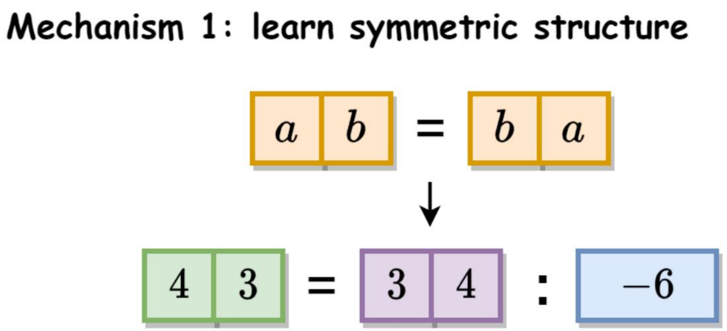
Padding

1 seen non-inferential composite anchors

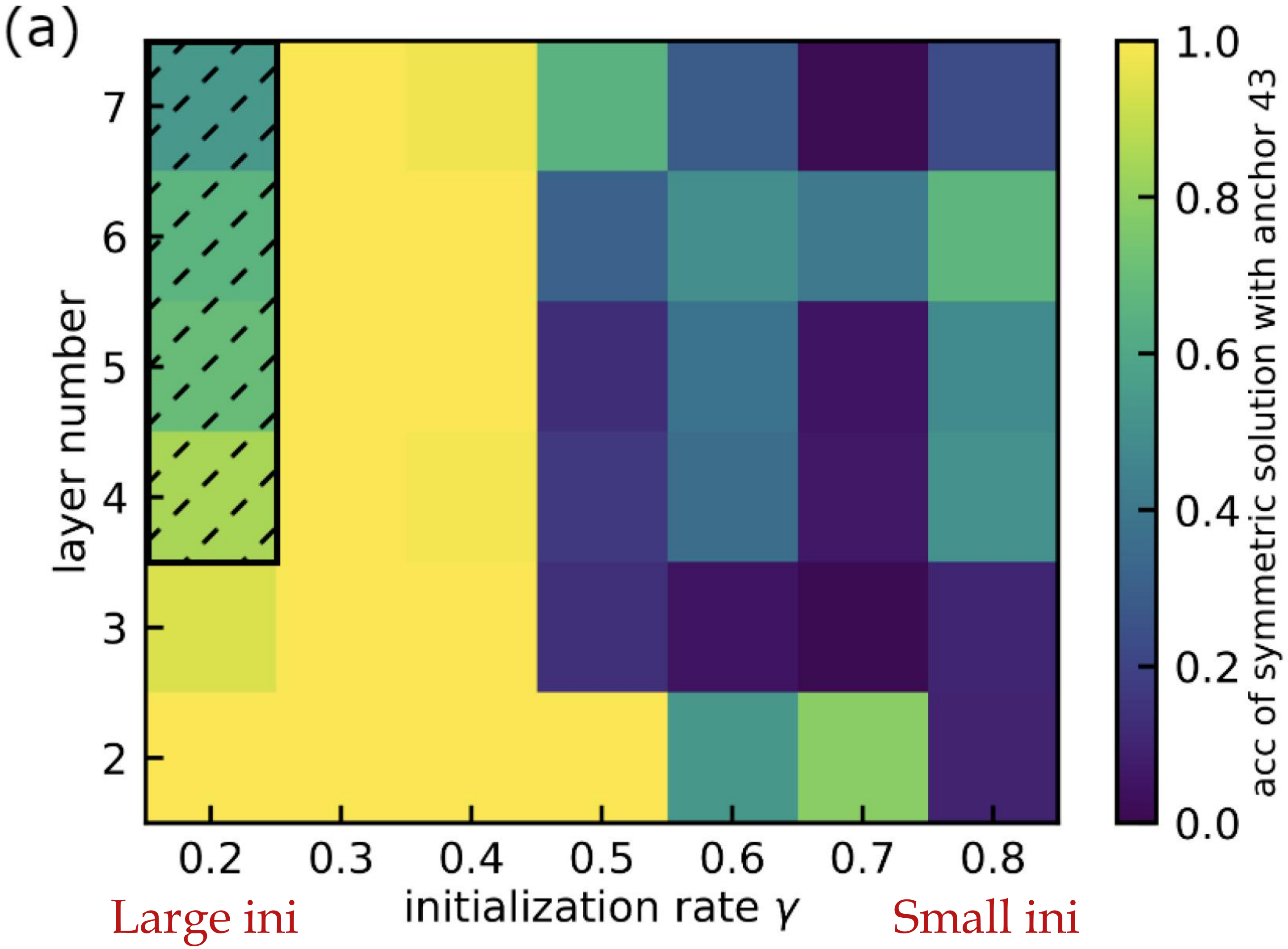
1 unseen composite anchor

Zhongwang Zhang, Pengxiao Lin, Zhiwei Wang, Yaoyu Zhang, Zhi-Qin John Xu*, Initialization is Critical to Whether Transformers Fit Composite Functions by Inference or Memorizing, NeurIPS 2024

Phase diagram of symmetric solution



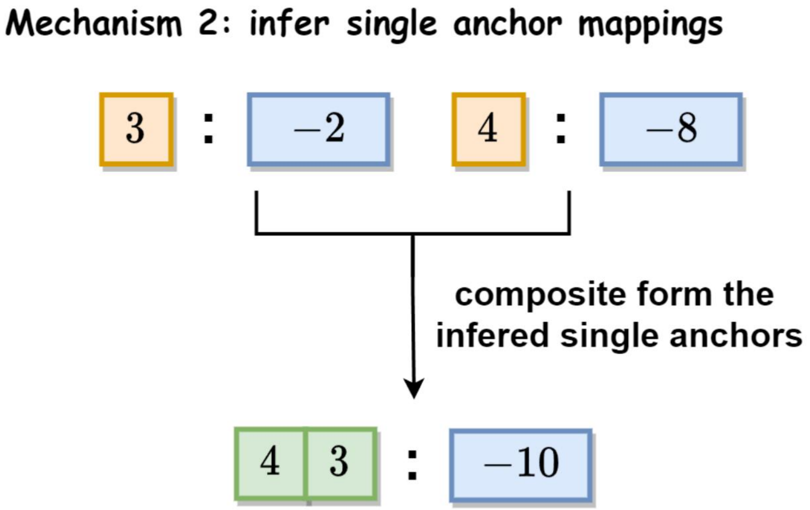
Initialization~$N\left(0,\frac{1}{d_{in}^{\gamma}}\right)$



▨ bad generalization on seen anchors (test accuracy < 90%)

Zhongwang Zhang, Pengxiao Lin, Zhiwei Wang, Yaoyu Zhang, Zhi-Qin John Xu*, Initialization is Critical to Whether Transformers Fit Composite Functions by Inference or Memorizing, NeurIPS 2024

Phase diagram of inferential solution



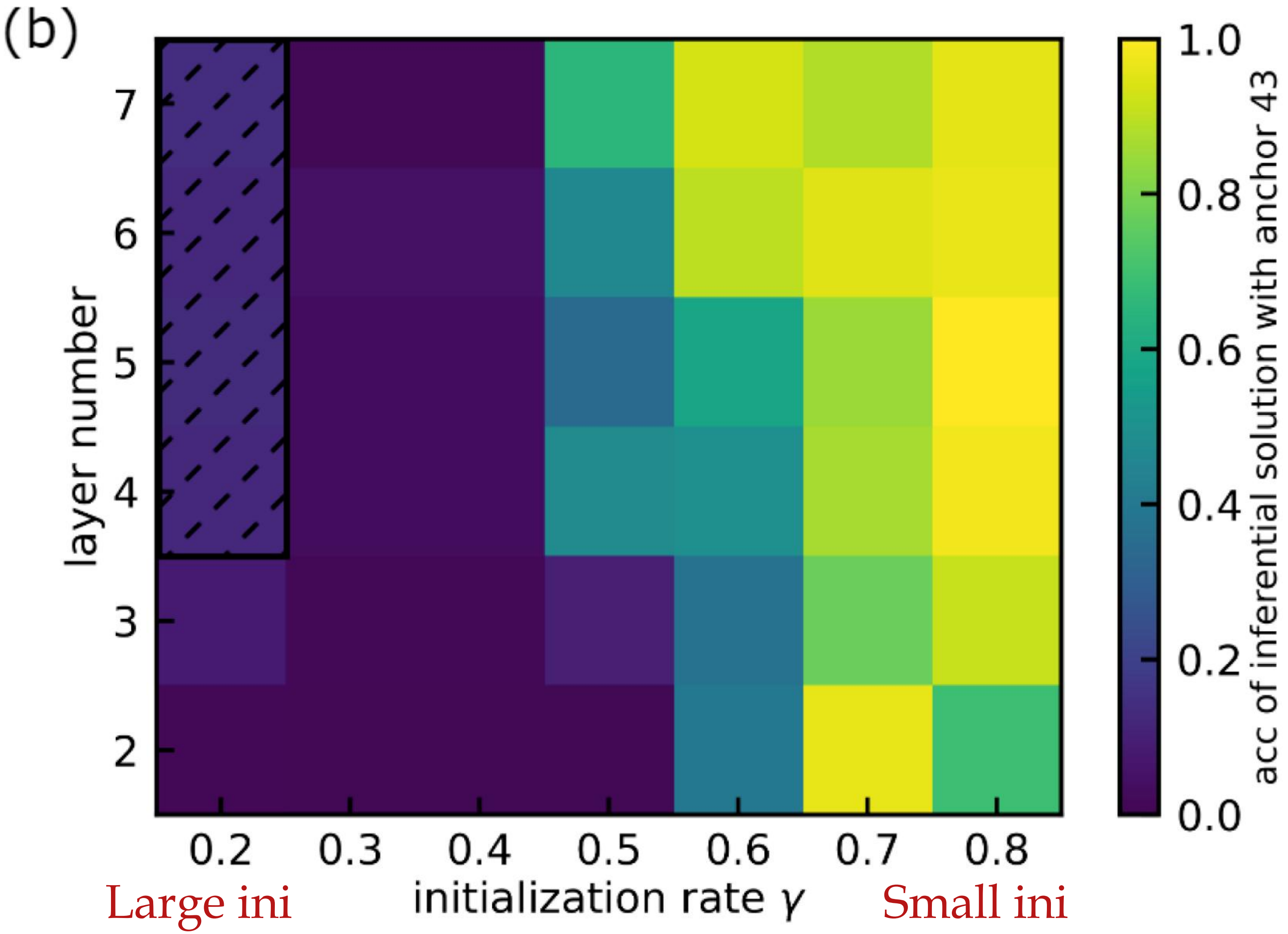
Mechanism 2: infer single anchor mappings

3 : -2 4 : -8

composite form the infered single anchors

4 3 : -10

$$\text{Initialization} \sim N\left(0, \frac{1}{d_{in}^{\gamma}}\right)$$

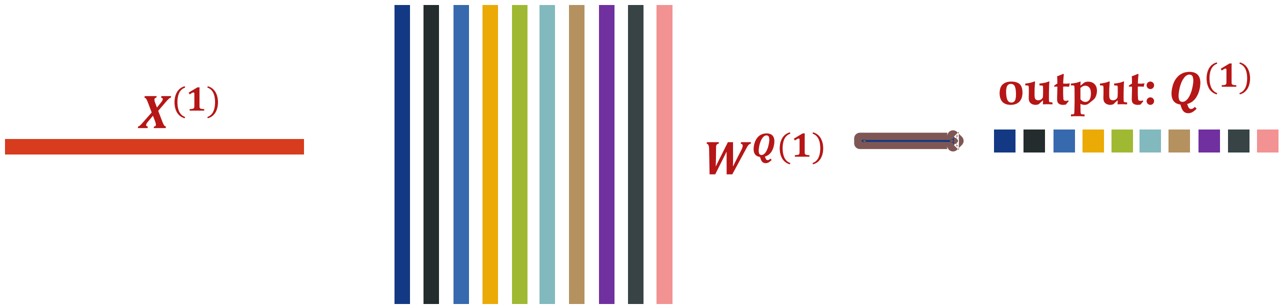


▨ bad generalization on seen anchors (test accuracy < 90%)

Zhongwang Zhang, Pengxiao Lin, Zhiwei Wang, Yaoyu Zhang, Zhi-Qin John Xu*, Initialization is Critical to Whether Transformers Fit Composite Functions by Inference or Memorizing, NeurIPS 2024

Condensation of $W^{Q(1)}$ by column

$$Q^{(l)} = X^{(l)} W^{Q(l)}$$



$X^{(1)}$

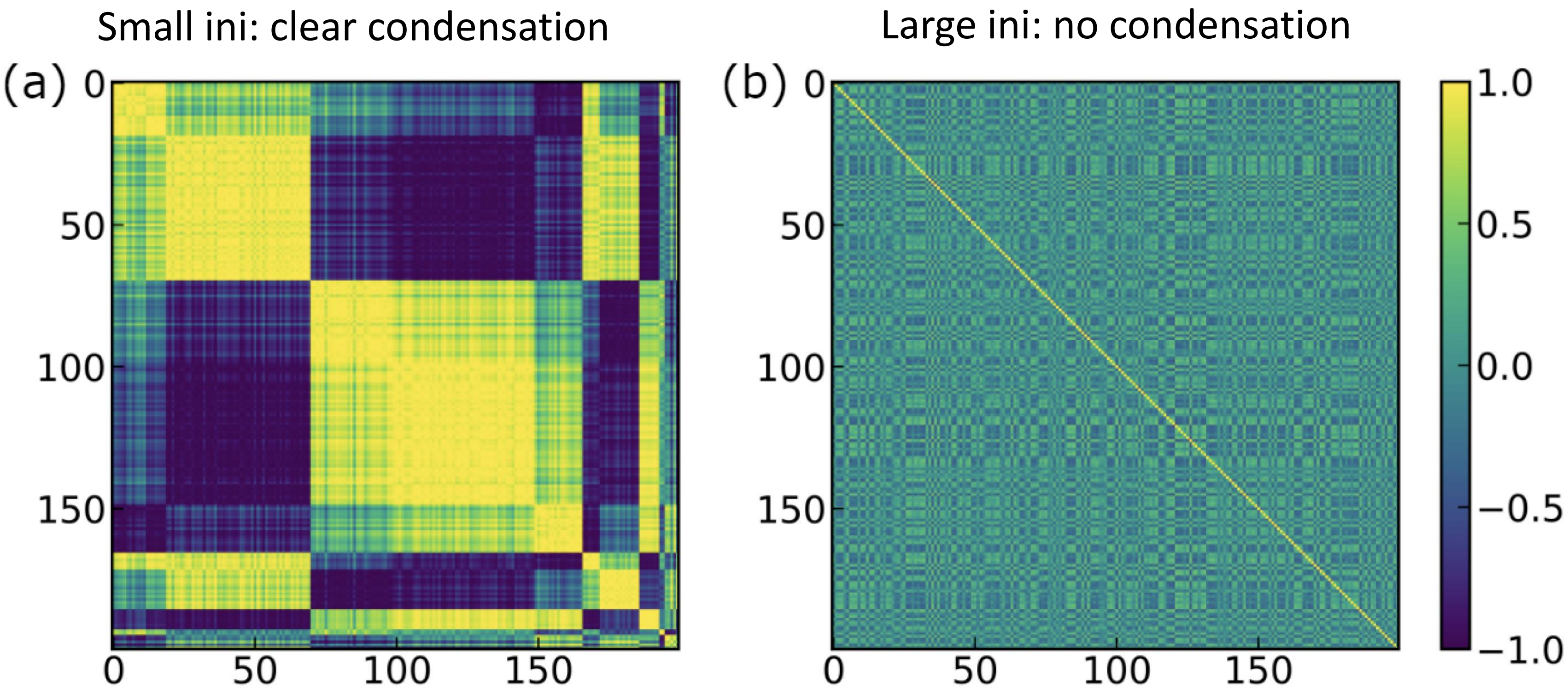
$W^{Q(1)}$

output: $Q^{(1)}$

Cosine Similarity b/w

Columns of $W^{Q(1)}$

Input weight of Q neurons



Small ini: clear condensation

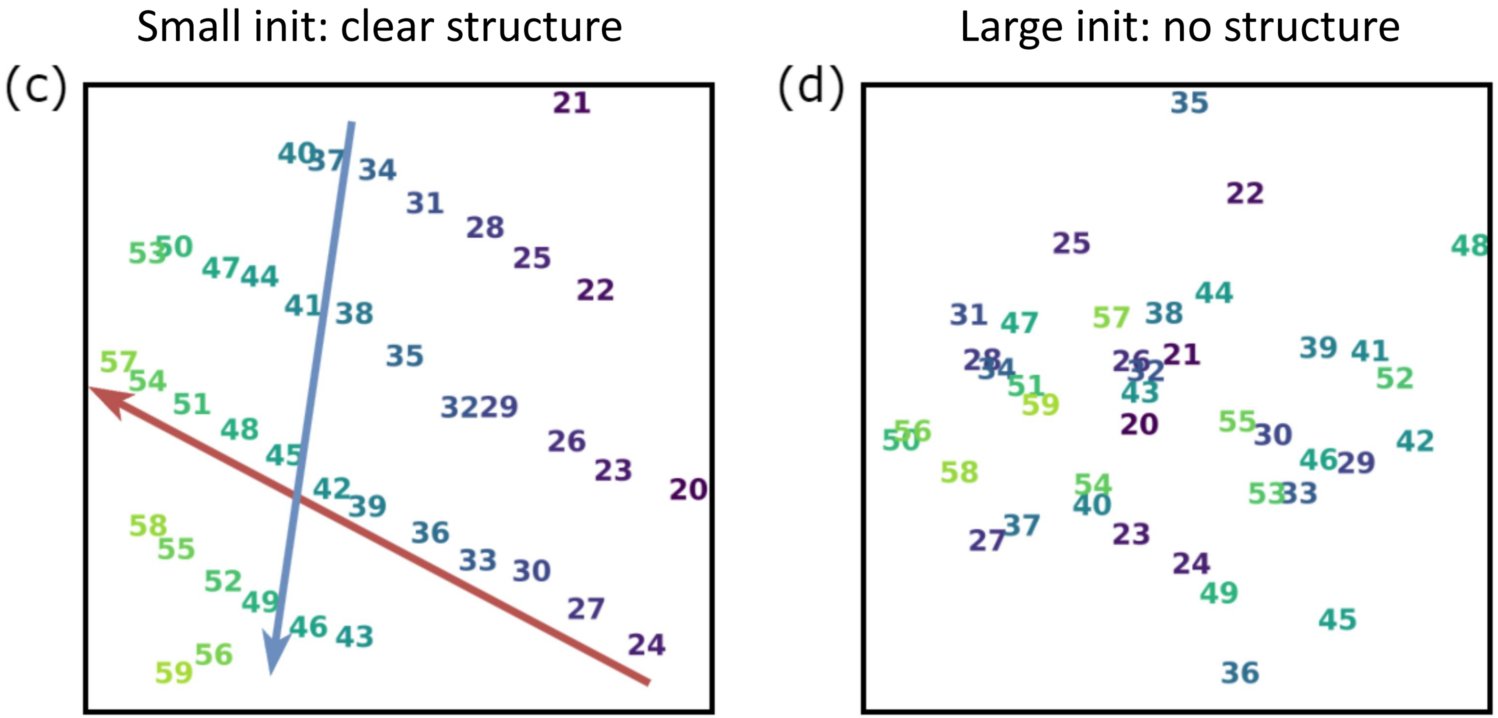
Large ini: no condensation

Zhongwang Zhang, Pengxiao Lin, Zhiwei Wang, Yaoyu Zhang, Zhi-Qin John Xu*, Initialization is Critical to Whether Transformers Fit Composite Functions by Inference or Memorizing, NeurIPS 2024

Structural Features of the Word Embedding

T-SNE

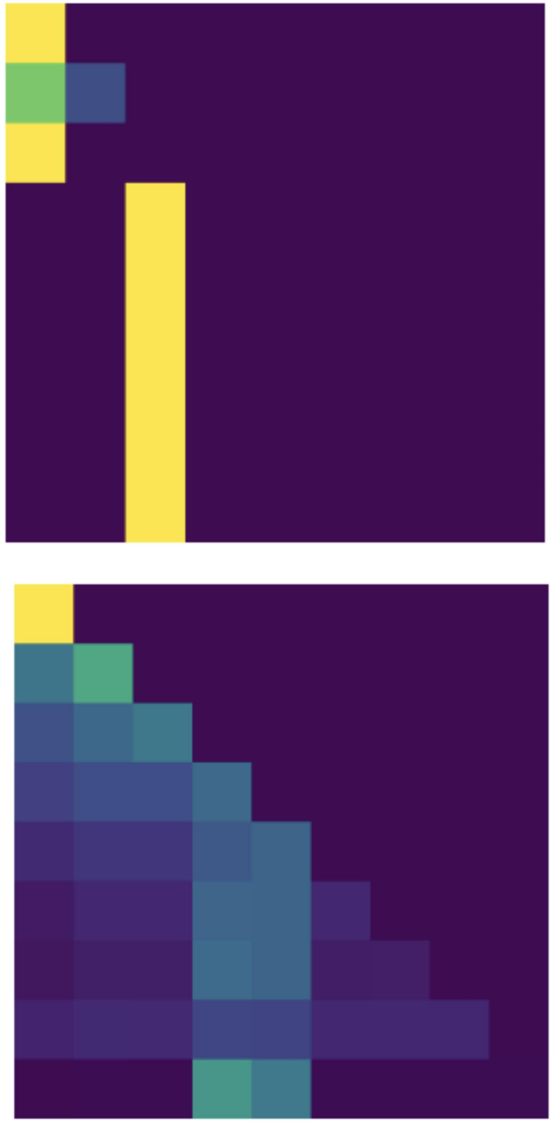
$$f_a(x) = \begin{cases} x+5, \; if \; a \; = 1 \\ x+1, \; if \; a \; = 2 \\ x-2, \; if \; a \; = 3 \\ x-8, \; if \; a \; = 4 \end{cases}$$

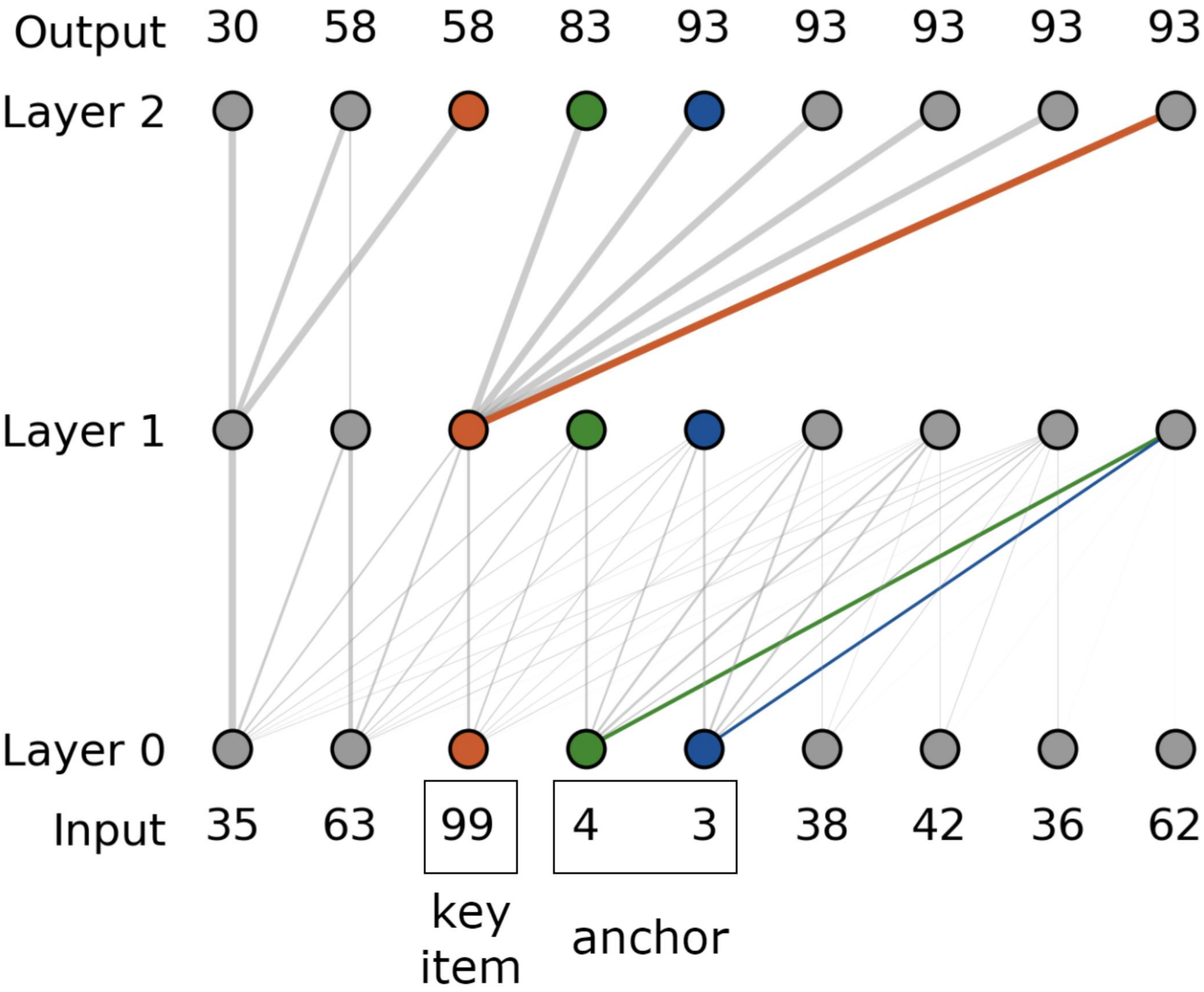

Small init: clear structure

Large init: no structure

Zhongwang Zhang, Pengxiao Lin, Zhiwei Wang, Yaoyu Zhang, Zhi-Qin John Xu*, Initialization is Critical to Whether Transformers Fit Composite Functions by Inference or Memorizing, NeurIPS 2024

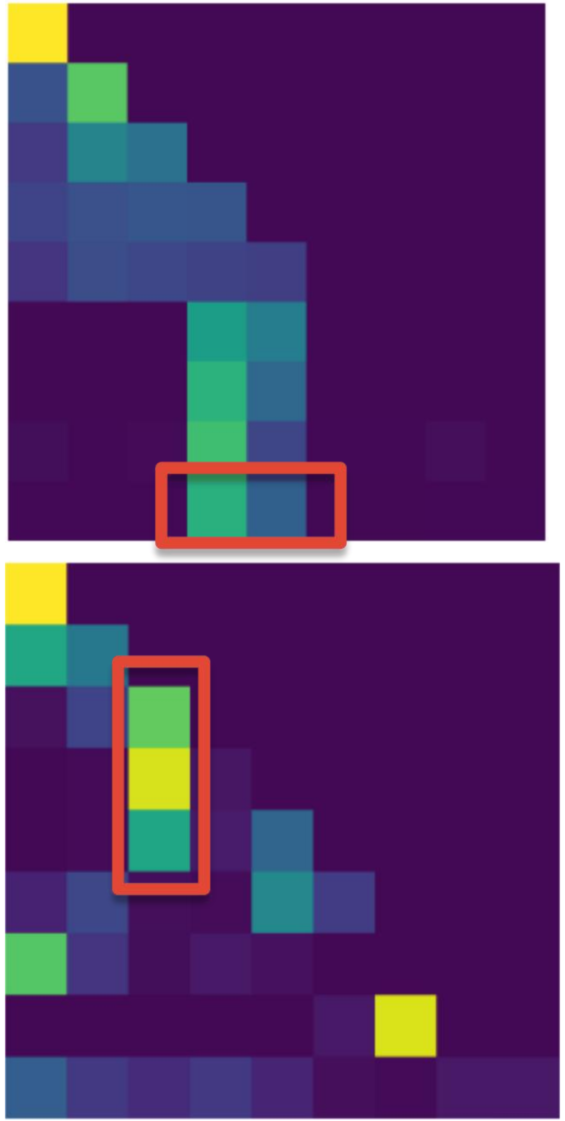
Symmetric solution

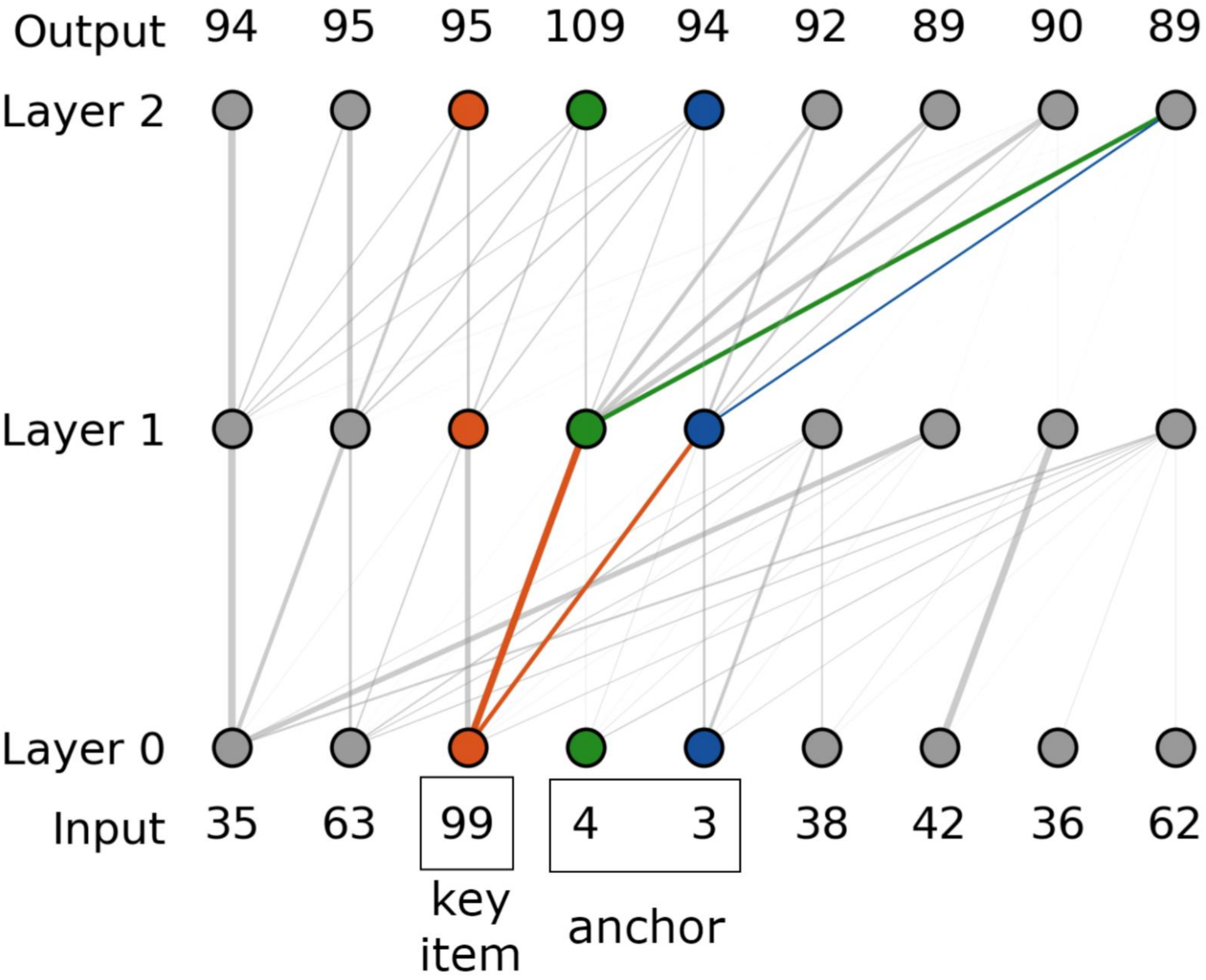



Output 30 58 58 83 93 93 93 93 93
Layer 2
Layer 1
Layer 0
Input 35 63 99 4 3 38 42 36 62
key item
anchor

Zhongwang Zhang, Pengxiao Lin, Zhiwei Wang, Yaoyu Zhang, Zhi-Qin John Xu*, Initialization is Critical to Whether Transformers Fit Composite Functions by Inference or Memorizing, NeurIPS 2024

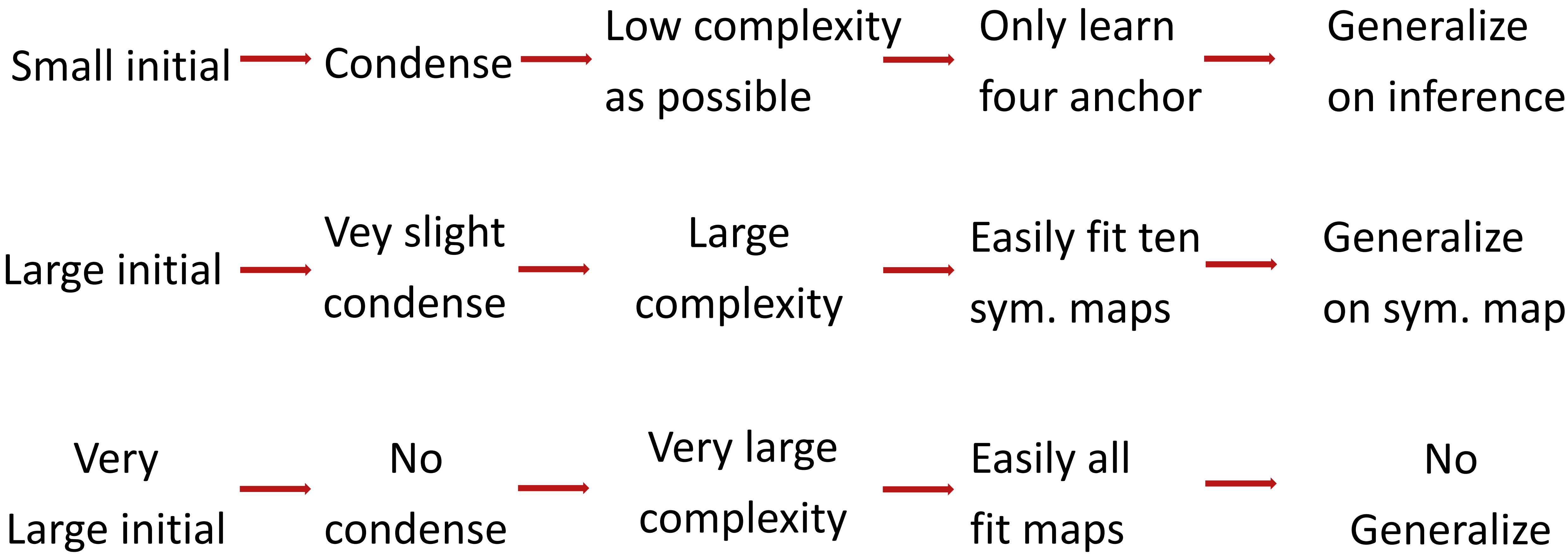
Inferential solution





Zhongwang Zhang, Pengxiao Lin, Zhiwei Wang, Yaoyu Zhang, Zhi-Qin John Xu*, Initialization is Critical to Whether Transformers Fit Composite Functions by Inference or Memorizing, NeurIPS 2024

Mechanisms underlying initialization effect



Small initial → Condense → Low complexity as possible → Only learn four anchor → Generalize on inference

Large initial → Vey slight condense → Large complexity → Easily fit ten sym. maps → Generalize on sym. map

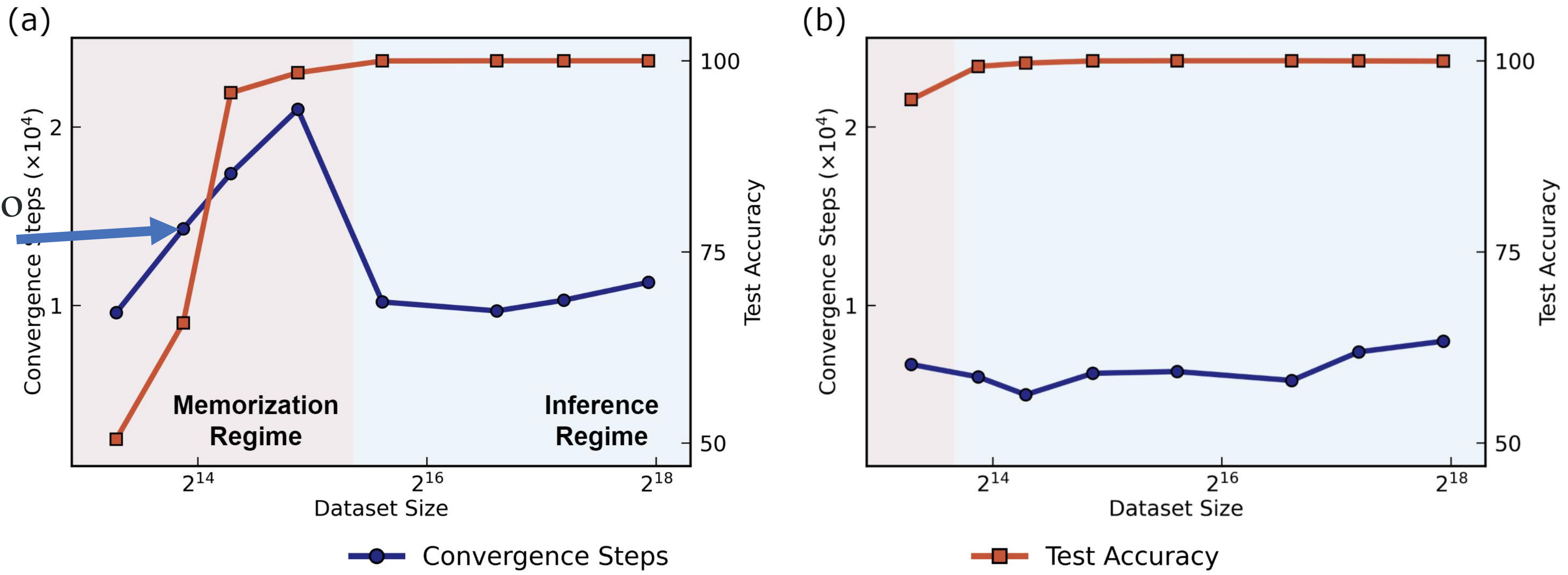
Very Large initial → No condense → Very large complexity → Easily all fit maps → No Generalize

Zhongwang Zhang, Pengxiao Lin, Zhiwei Wang, Yaoyu Zhang, Zhi-Qin John Xu*, Initialization is Critical to Whether Transformers Fit Composite Functions by Inference or Memorizing, NeurIPS 2024

Other realistic reasoning dataset

Large initial → Large complexity → Memorization to inference

Small initial → Low complexity as possible → Inference



More data, more steps to memorize

Zhongwang Zhang, Pengxiao Lin, Zhiwei Wang, Yaoyu Zhang, Zhi-Qin John Xu*, Initialization is Critical to Whether Transformers Fit Composite Functions by Inference or Memorizing, NeurIPS 2024

上海交通大學

SHANGHAI JIAO TONG UNIVERSITY