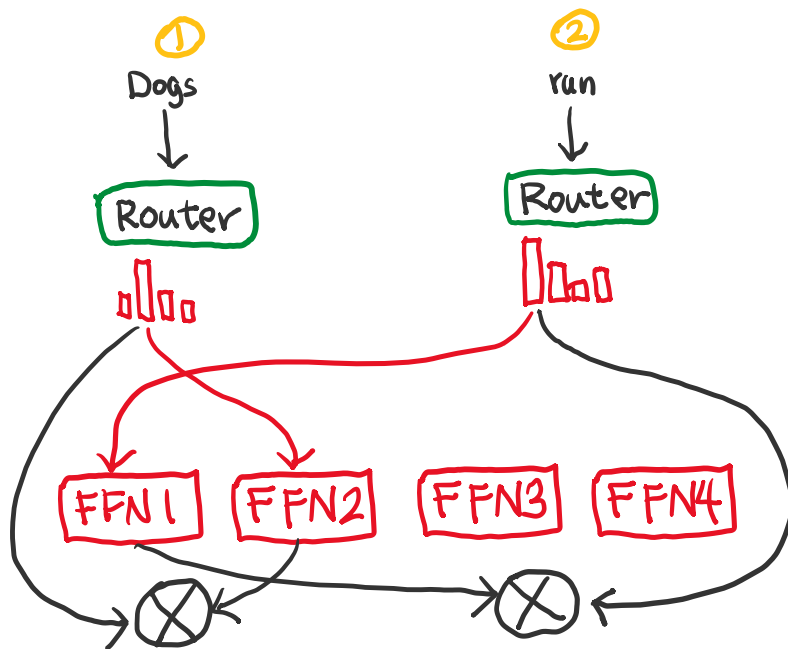
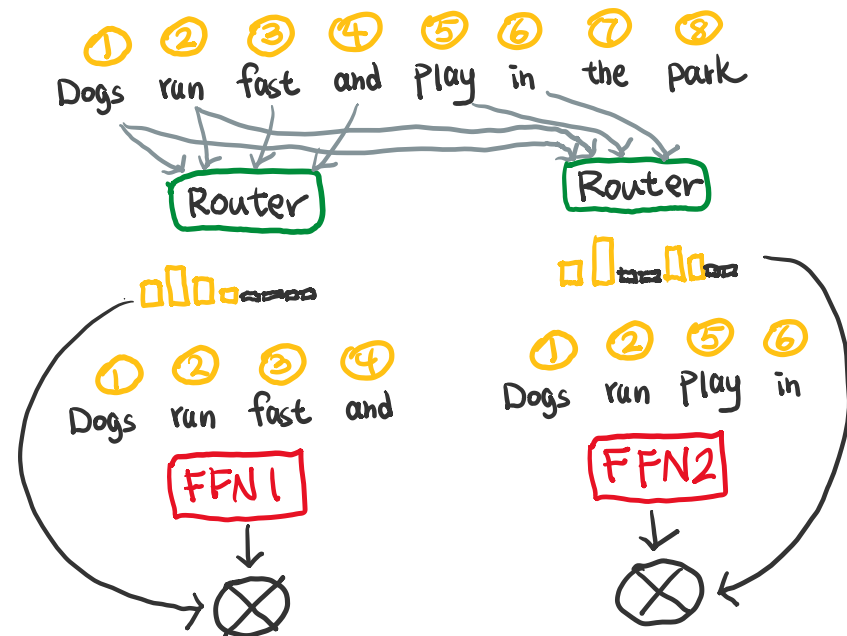


Expert Choice Mixture of Experts

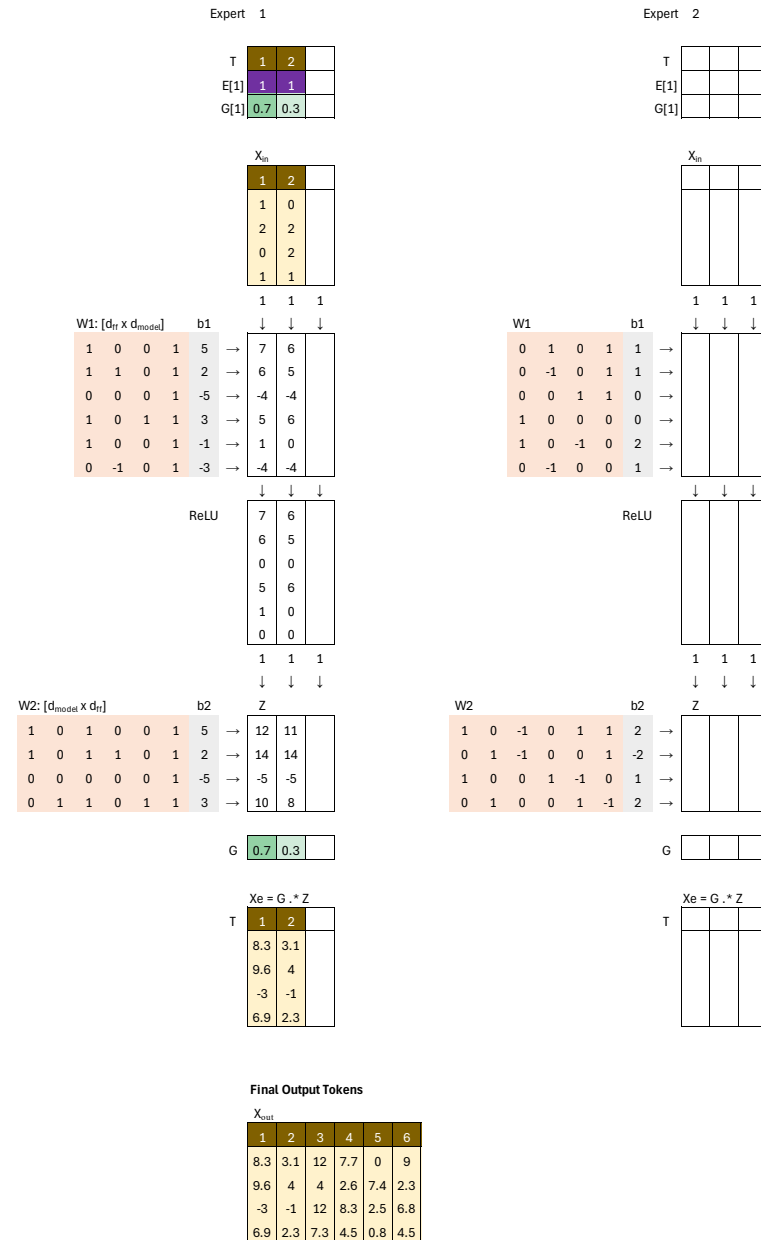
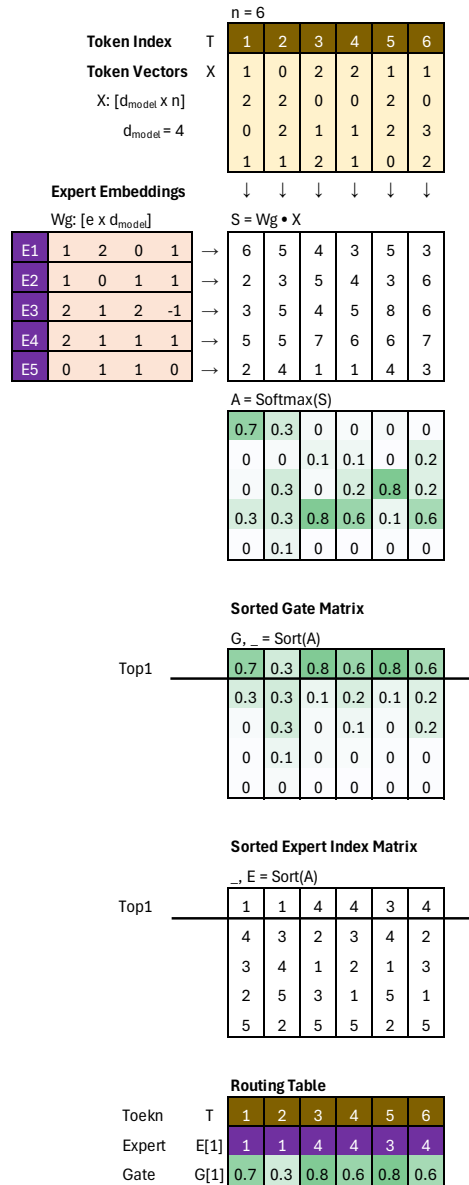
MoE



Expert Choice MoE



Mixture of Experts



Expert 3

T	5		
E[1]	3		
G[1]	0.8		

X_{in}	5		
	1		
	2		
	2		
	0		

W1		b1			
	1	1	0	1	1
	1	1	0	1	0
	0	0	1	0	2
	1	0	1	0	3
	1	0	0	1	1
	0	1	0	0	-4

ReLU	4		
	3		
	4		
	6		
	2		
	0		

W2		b2			
	1	0	1	-1	0
	1	0	-1	1	0
	0	0	0	0	1
	0	1	-1	0	0

G	0.8		
---	-----	--	--

$X_e = G .* Z$	5		
T	0		
	7.4		
	2.5		
	0.8		

Expert 4

T	3	4	6
E[1]	4	4	4
G[1]	0.8	0.6	0.6

X_{in}	3	4	6
	2	2	1
	0	0	0
	1	1	3
	2	1	2

W1		b1			
	1	0	0	1	1
	1	1	0	0	2
	0	1	0	1	0
	0	0	1	1	3
	1	0	0	1	0
	0	-1	0	1	-3

ReLU	5	4	4
	4	4	3
	2	1	2
	6	5	8
	4	3	3
	0	0	0

W2		b2			
	1	0	1	1	0
	1	0	0	0	0
	1	1	0	0	1
	0	0	1	0	1

G	0.8	0.6	0.6
---	-----	-----	-----

$X_e = G .* Z$	3	4	6
T	12	7.7	9
	4	2.6	2.3
	12	8.3	6.8
	7.3	4.5	4.5

Expert 5

T			
E[1]			
G[1]			

X_{in}			

W1		b1			
	1	0	0	1	1
	1	1	0	1	2
	0	0	0	0	0
	1	0	1	0	1
	1	0	0	1	2
	0	-1	0	1	-3

ReLU			

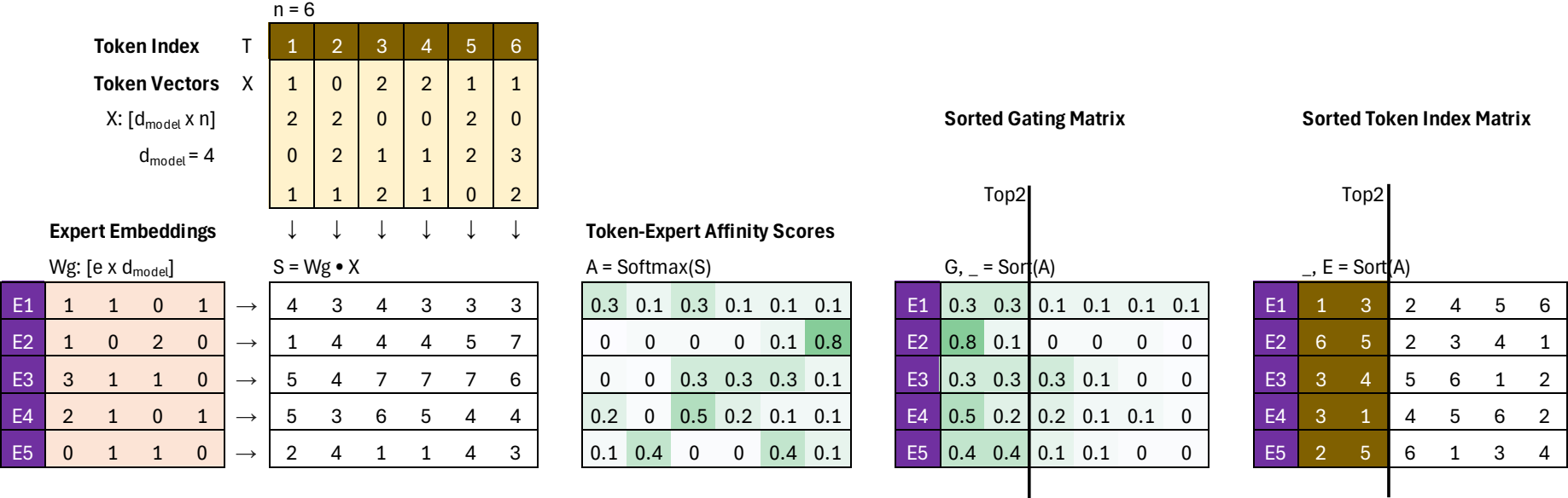
W2		b2			
	1	0	1	0	0
	1	0	0	0	1
	1	-1	0	1	1
	0	1	0	0	1

G			
---	--	--	--

$X_e = G .* Z$			
T			

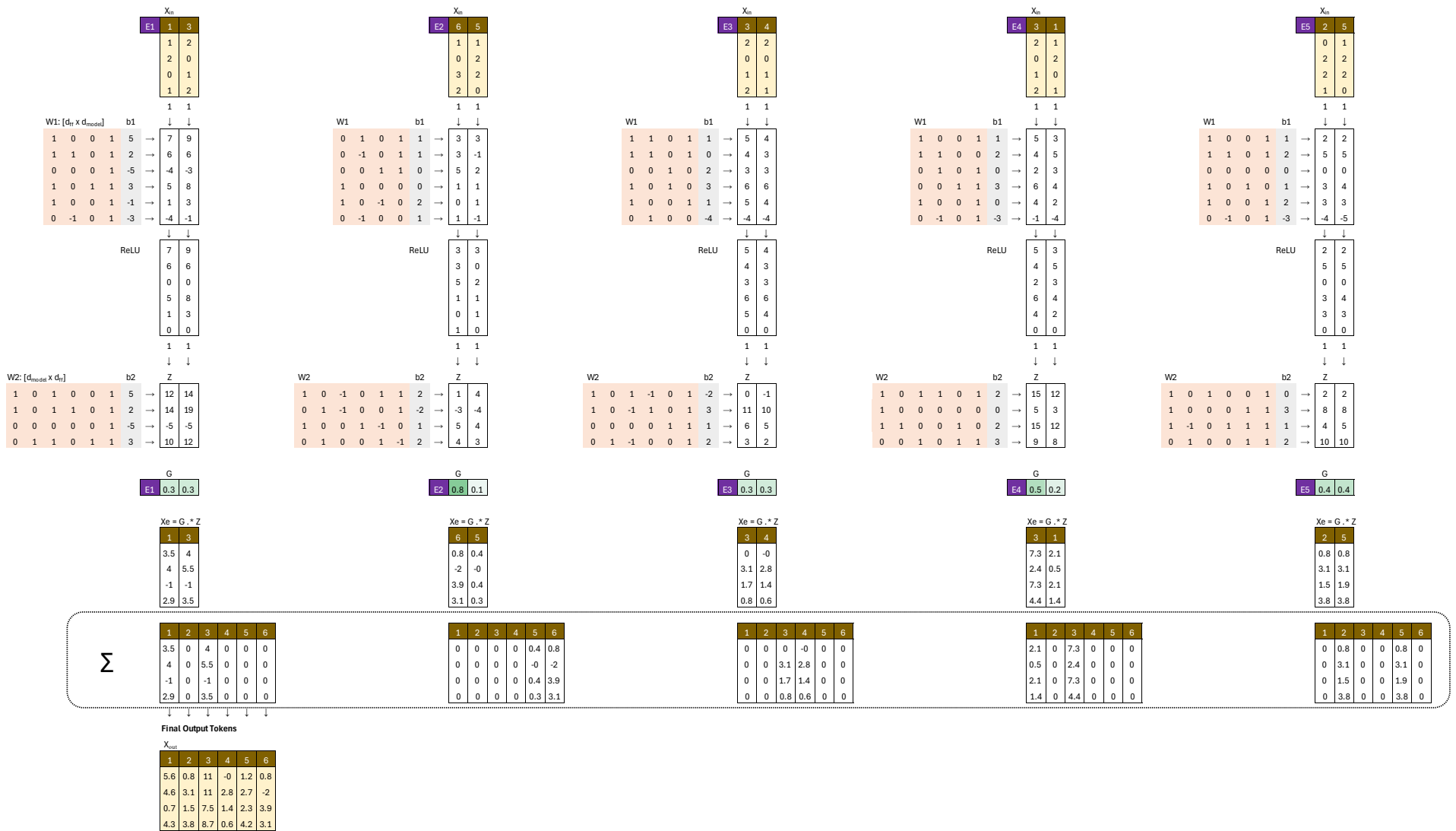
Expert-Choice Mixture of Experts

Number of Experts (e)	5
Model Dimensions (d_{model})	4
Sequence Length (n)	6
Top K	2
FFN Dimensions (d_{ff})	6



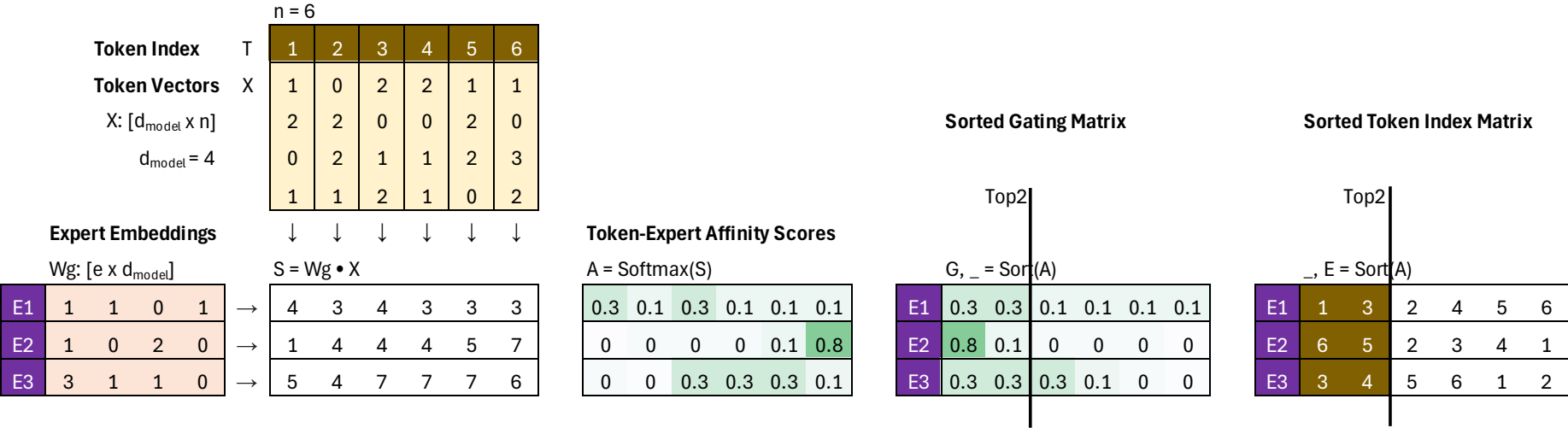
Sorted Token Index Matrix

$_, E = \text{Sort}(A)$



Expert-Choice Mixture of Experts

Number of Experts (e)	3
Model Dimensions (d_{model})	4
Sequence Length (n)	6
Top K	2
FFN Dimensions (d_{ff})	6



Sorted Gating Matrix

G, _ = Sort(A)

E1

0.3

0.3

0.1

0.1

0.1

0.1

E2

0.8

0.1

0

0

0

0

E3

0.3

0.3

0.3

0.1

0

0

Sorted Token Index Matrix

_ , E = Sort(A)

E1

1

3

2

4

5

6

E2

6

5

2

3

4

1

E3

3

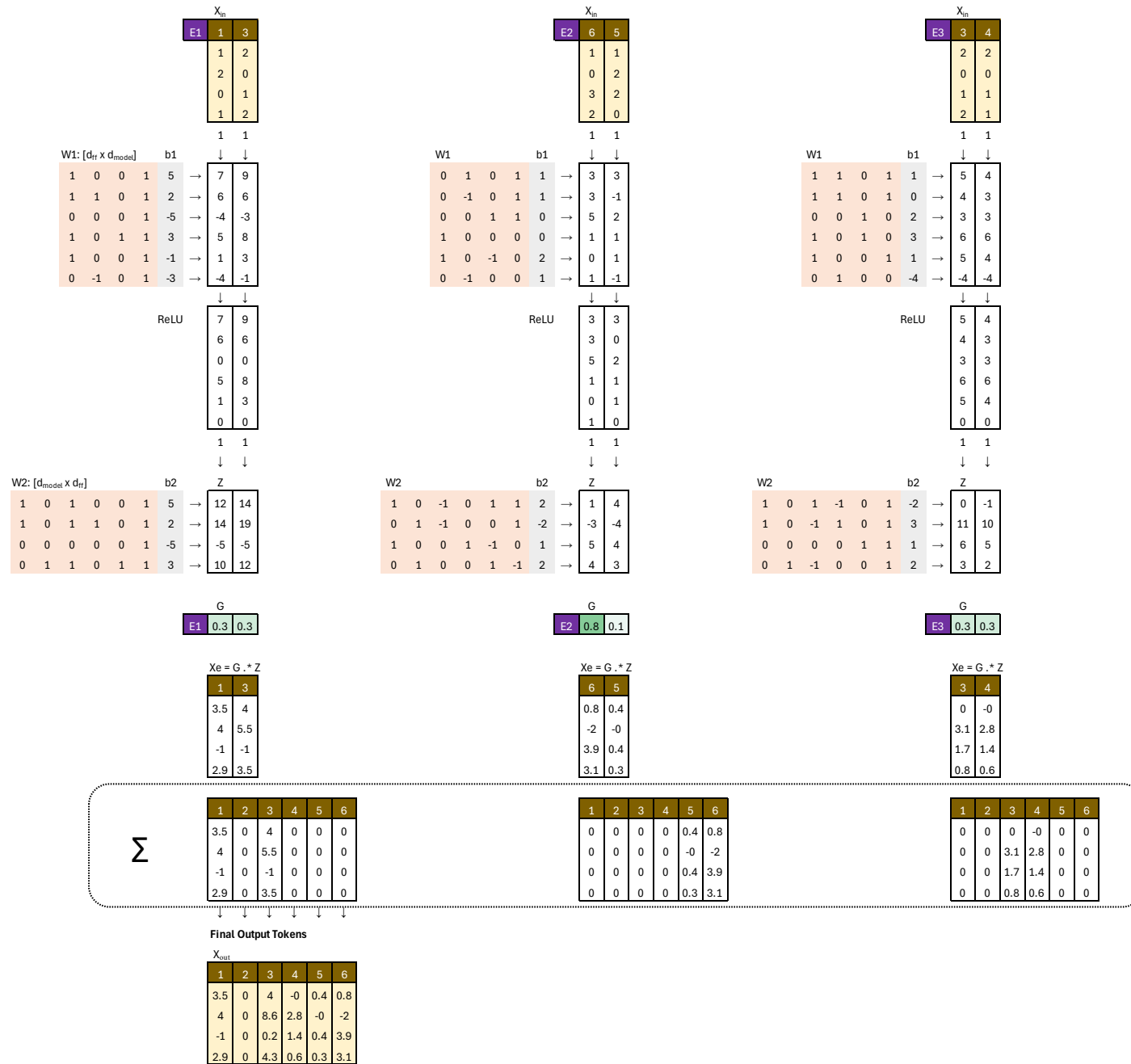
4

5

6

1

2



Expert-Choice Mixture of Experts

Number of Experts (e)	5
Model Dimensions (d_{model})	4
Sequence Length (n)	6
Top K	3
FFN Dimensions (d_{ff})	6

