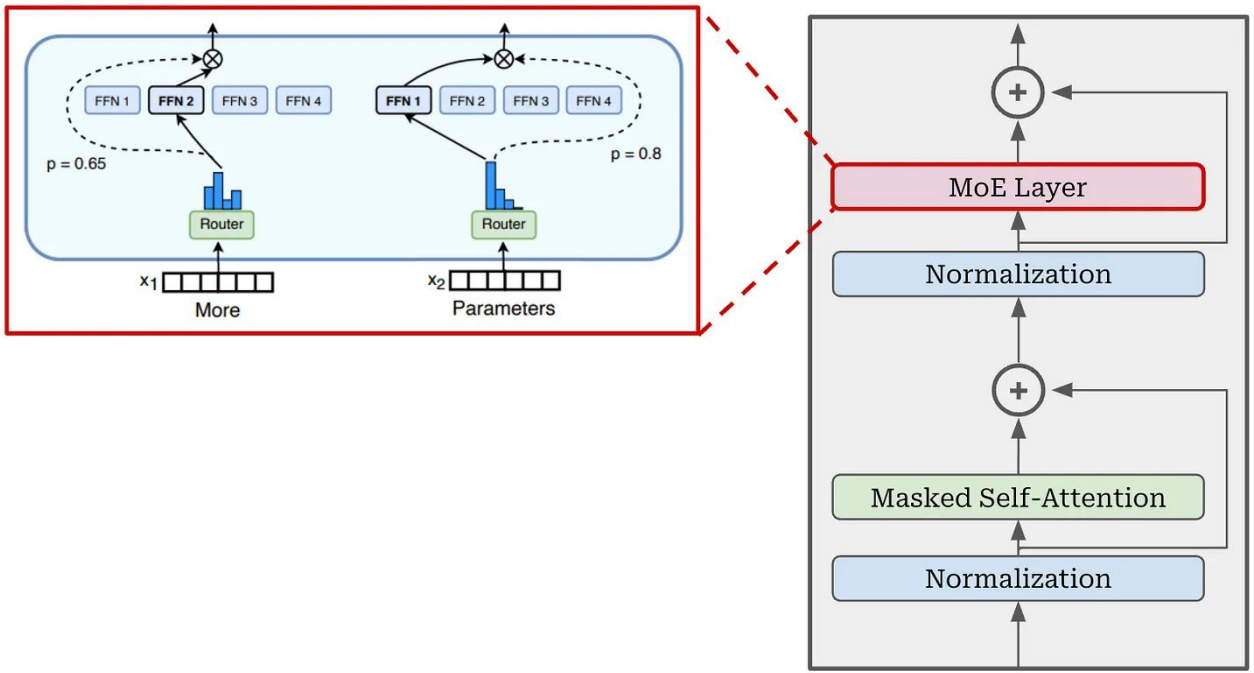
Token routing inside MoEs

Hugging Face

MoE Layer



Source

Problem Statement: *A toy problem*

- 4 Tokens
- 4 Experts
- 2 Slots per Expert

Which *token* routes to which *expert* and in what *slot*?

TOKENS

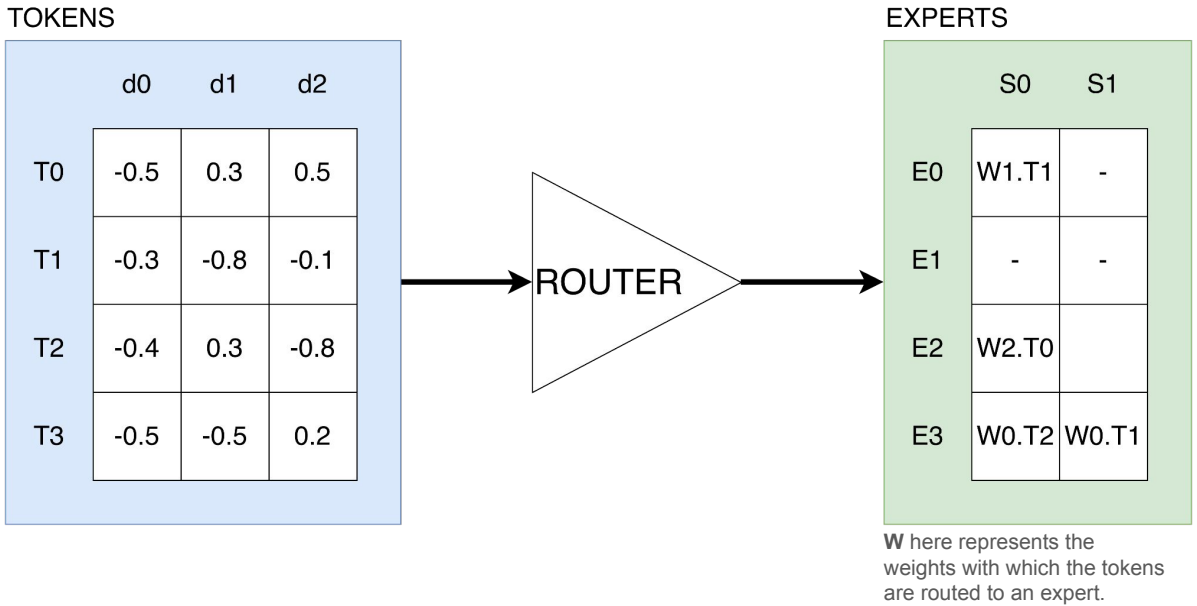
	d0	d1	d2
T0	-0.5	0.3	0.5
T1	-0.3	-0.8	-0.1
T2	-0.4	0.3	-0.8
T3	-0.5	-0.5	0.2

EXPERTS

	S0	S1
E0		
E1		
E2		
E3		

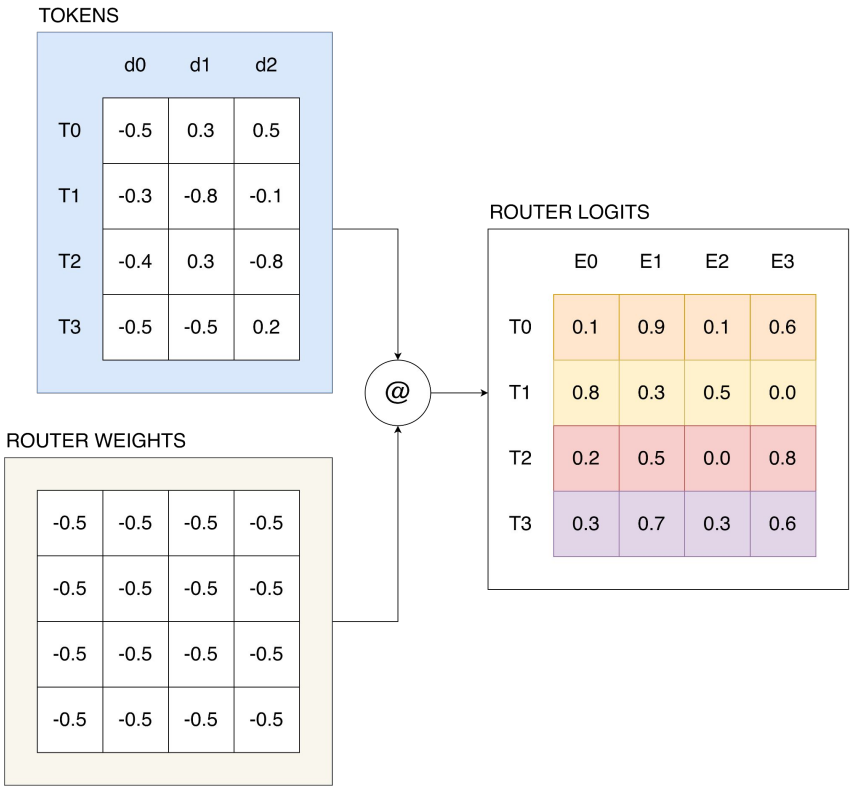
Our Job: To figure out how the router works 🕵️

- 4 Tokens (**T**)
- 4 Experts (**E**)
- 2 Slots per Expert (**S**)
- Routing Algorithm

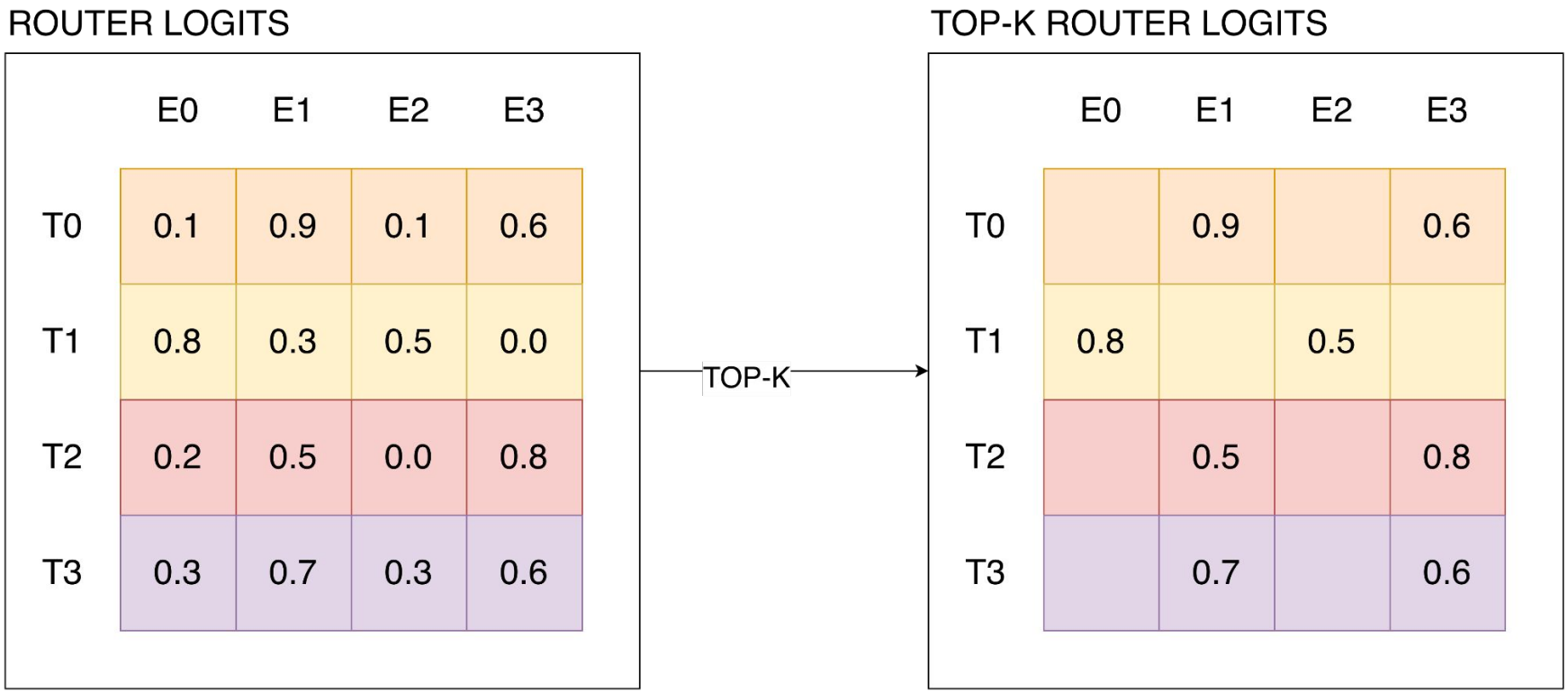


Token Weights

Compute Router Logits



Top-K Expert Section per Token



TOP-K Selection

ROUTER LOGITS

	E0	E1	E2	E3
T0	0.1	0.9	0.1	0.6
T1	0.8	0.3	0.5	0.0
T2	0.2	0.5	0.0	0.8
T3	0.3	0.7	0.3	0.6

TOP-K

TOP-K ROUTER LOGITS

	E0	E1	E2	E3
T0		0.9		0.6
T1	0.8		0.5	
T2		0.5		0.8
T3		0.7		0.6

TOP-2 LOGITS	
0.9	0.6
0.8	0.5
0.8	0.5
0.7	0.6

CHOSEN EXPERTS	
E1	E3
E0	E2
E3	E1
E1	E3

Normalize logits

TOP-K ROUTER LOGITS SCATTERED with -INF

	E0	E1	E2	E3
T0	-inf	0.9	-inf	0.6
T1	0.8	-inf	0.5	-inf
T2	-inf	0.5	-inf	0.8
T3	-inf	0.7	-inf	0.6

SOFTMAX

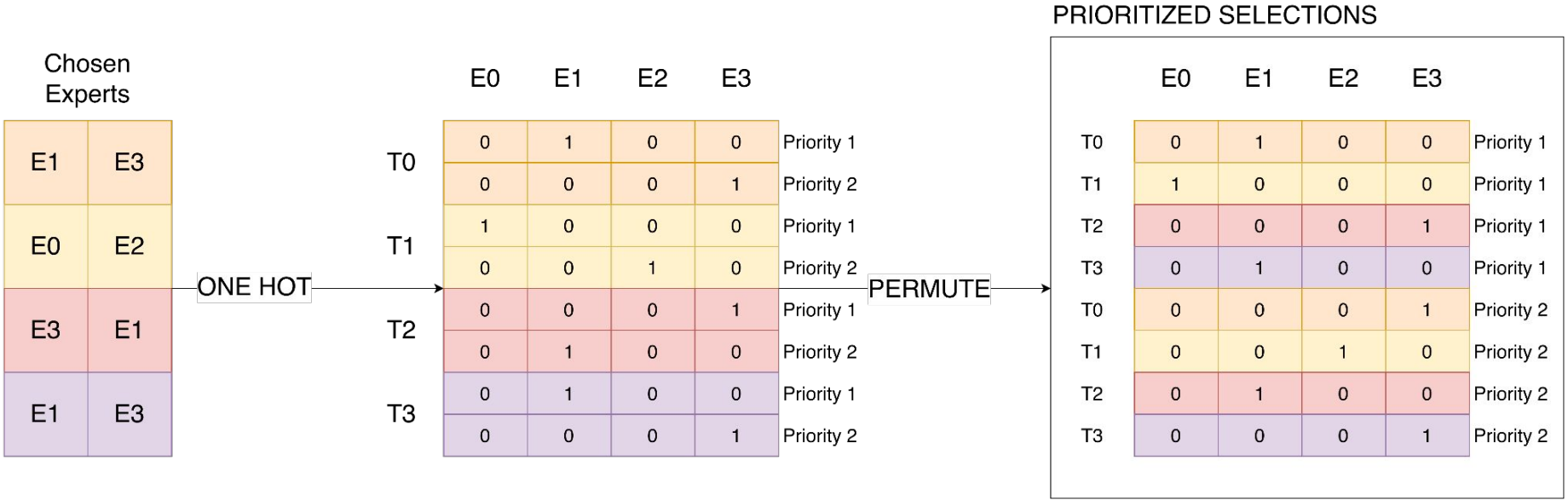
NORMALIZED TOKEN WEIGHTS

	E0	E1	E2	E3
T0	0.0	**0.5**	0.0	**0.4**
T1	**0.5**	0.0	**0.4**	0.0
T2	0.0	**0.4**	0.0	**0.5**
T3	0.0	**0.5**	0.0	**0.4**

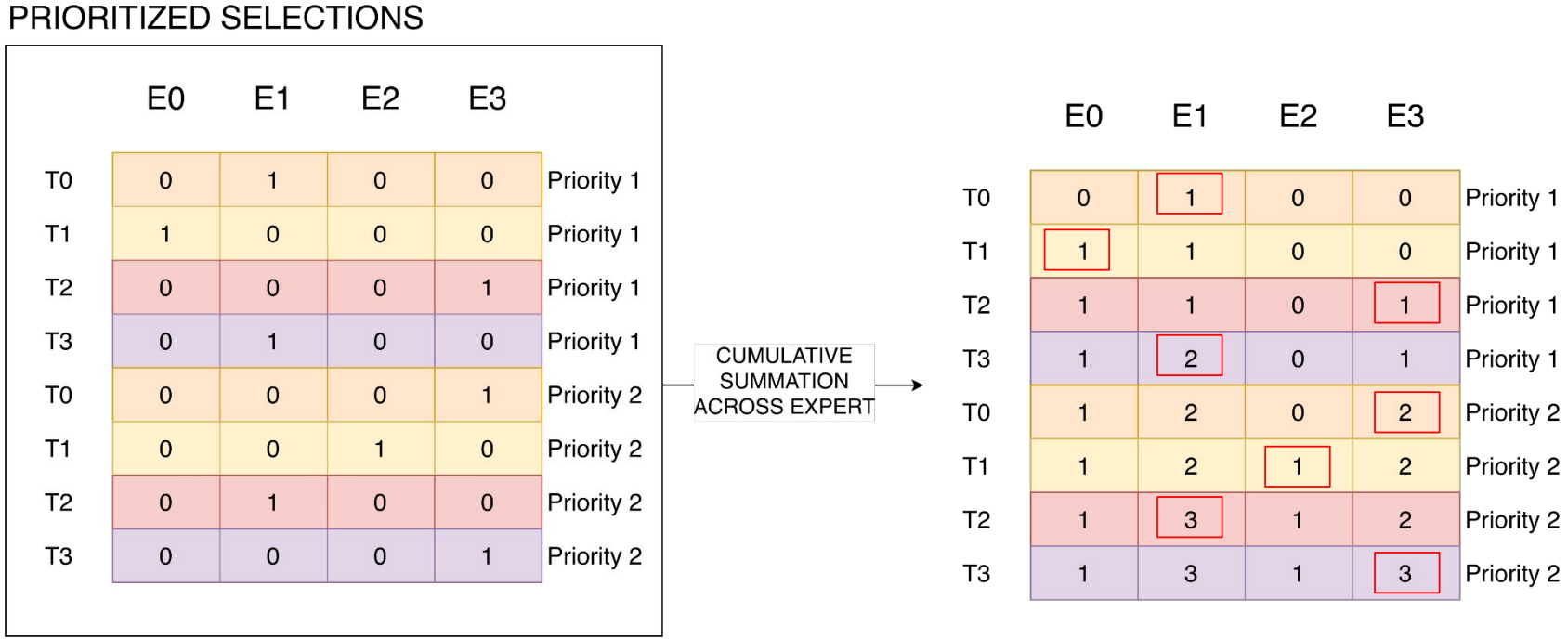
For the sake of *simplicity* we have **dropped** some decimal points as a consequence of which the normalized weights don't add up to 1. You can follow the code for the correct numbers.

Slot Selection

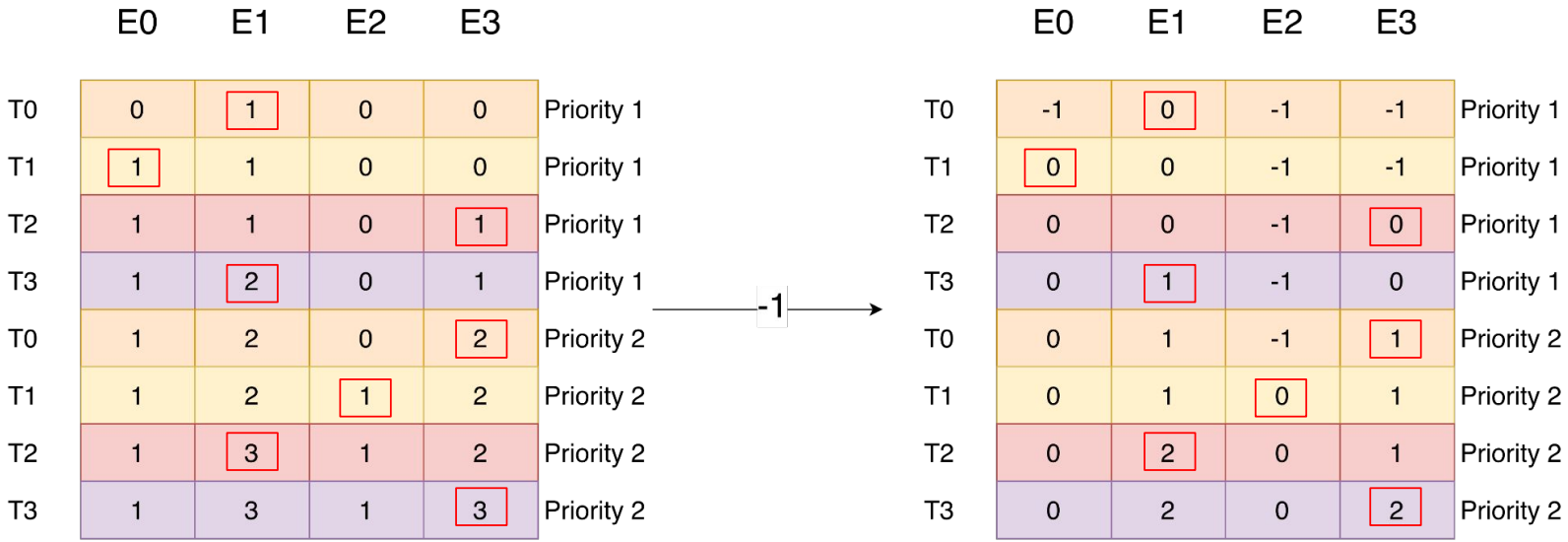
Prioritise the selection



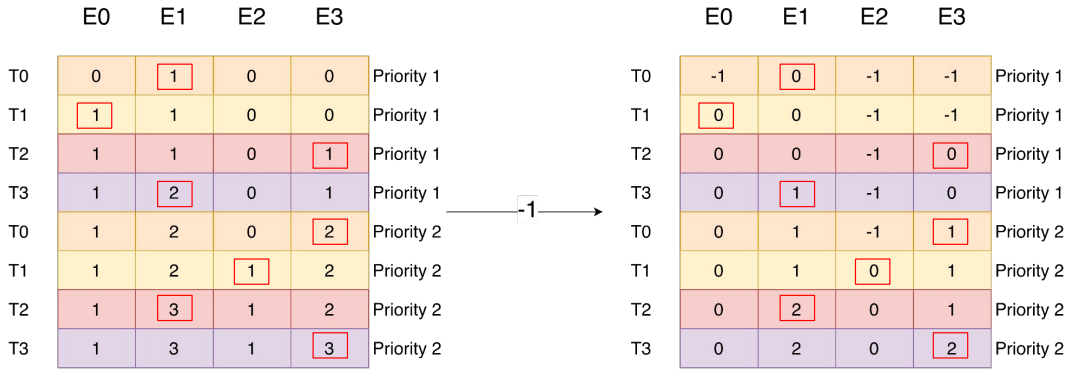
How many tokens have selected each expert *so far*?



Select Slots



Read Selected Slots

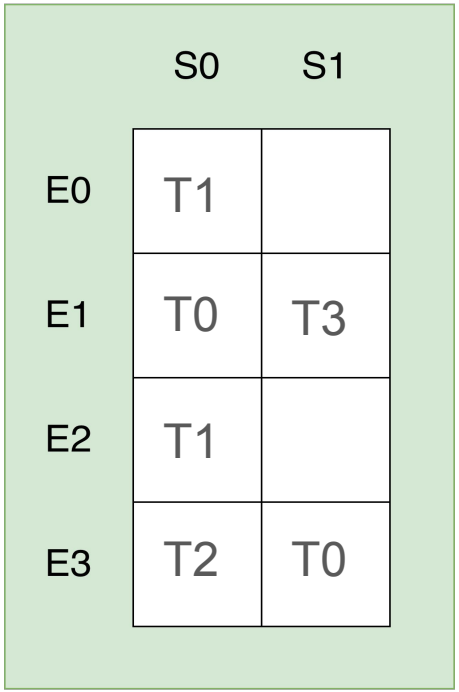


	Position 0	Position 1	Position 2
E0	Token 1	—	—
E1	Token 0	Token 3	Token 2
E2	Token 1	—	—
E3	Token 2	Token 0	Token 3

Problem with Slot Selection 😭

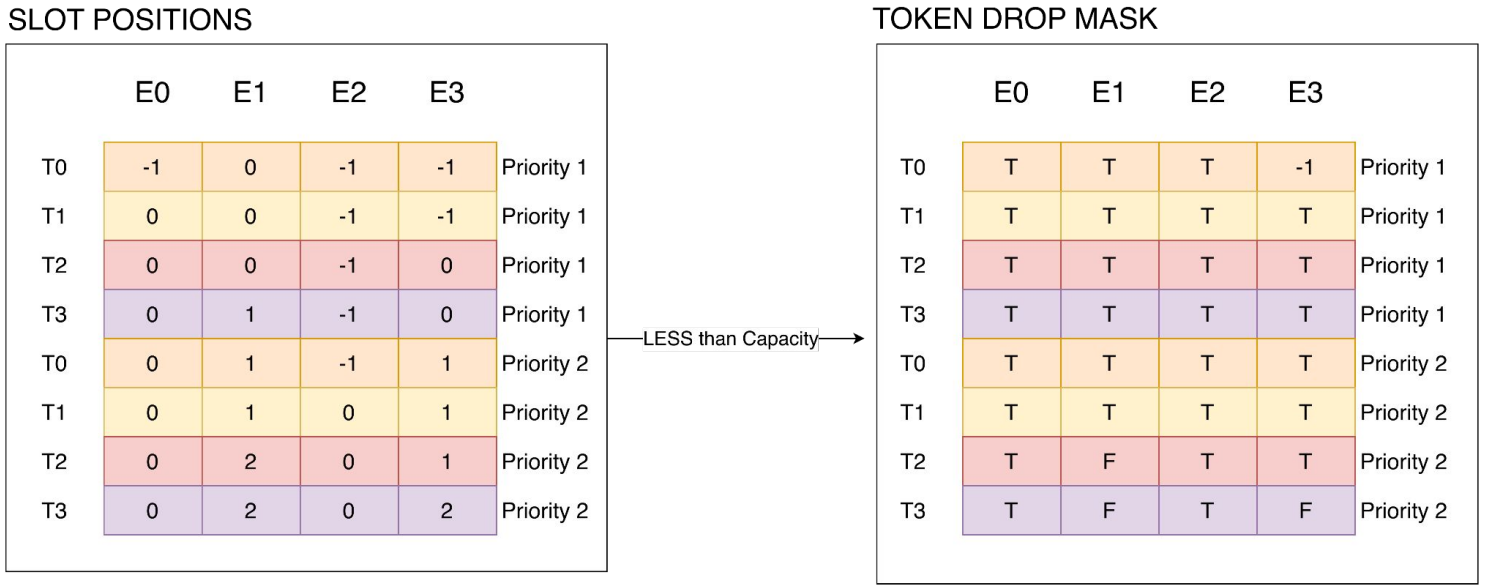
Place the tokens on respective slots

EXPERTS



	Position 0	Position 1	Position 2
E0	Token 1	—	—
E1	Token 0	Token 3	Token 2
E2	Token 1	—	—
E3	Token 2	Token 0	Token 3

Valid assignment mask



Drop tokens

TOKEN DROP MASK

	E0	E1	E2	E3	
T0	T	T	T	-1	Priority 1
T1	T	T	T	T	Priority 1
T2	T	T	T	T	Priority 1
T3	T	T	T	T	Priority 1
T0	T	T	T	T	Priority 2
T1	T	T	T	T	Priority 2
T2	T	F	T	T	Priority 2
T3	T	F	T	F	Priority 2

PRIORITIZED SELECTIONS

	E0	E1	E2	E3	
T0	0	1	0	0	Priority 1
T1	1	0	0	0	Priority 1
T2	0	0	0	1	Priority 1
T3	0	1	0	0	Priority 1
T0	0	0	0	1	Priority 2
T1	0	0	1	0	Priority 2
T2	0	1	0	0	Priority 2
T3	0	0	0	1	Priority 2

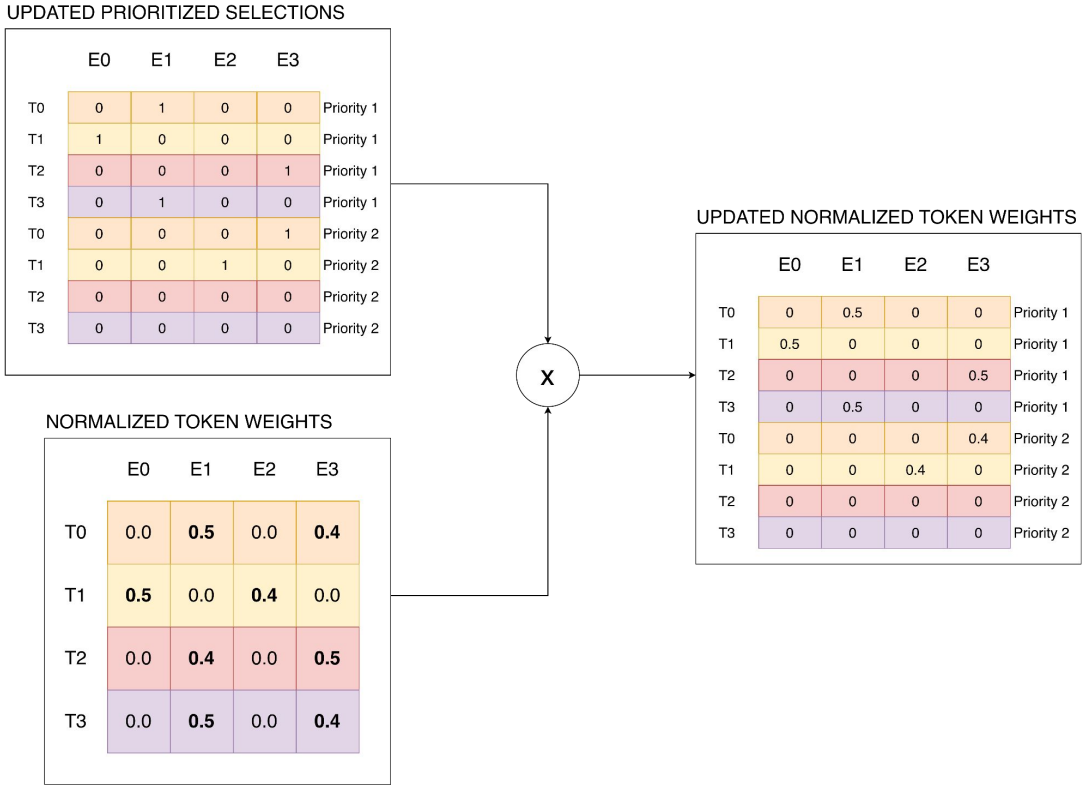
X

UPDATED PRIORITIZED SELECTIONS

	E0	E1	E2	E3	
T0	0	1	0	0	Priority 1
T1	1	0	0	0	Priority 1
T2	0	0	0	1	Priority 1
T3	0	1	0	0	Priority 1
T0	0	0	0	1	Priority 2
T1	0	0	1	0	Priority 2
T2	0	0	0	0	Priority 2
T3	0	0	0	0	Priority 2

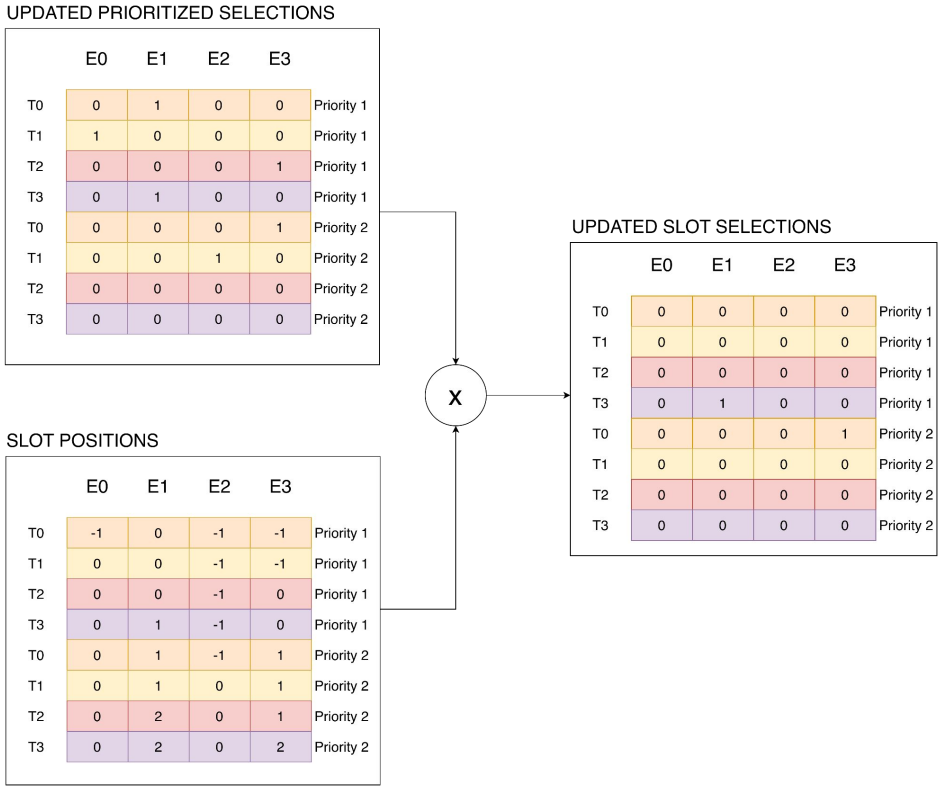
	Position 0	Position 1	Position 2
E0	Token 1	—	—
E1	Token 0	Token 3	~~Token 2~~
E2	Token 1	—	—
E3	Token 2	Token 0	~~Token 3~~

Update normalized token weights

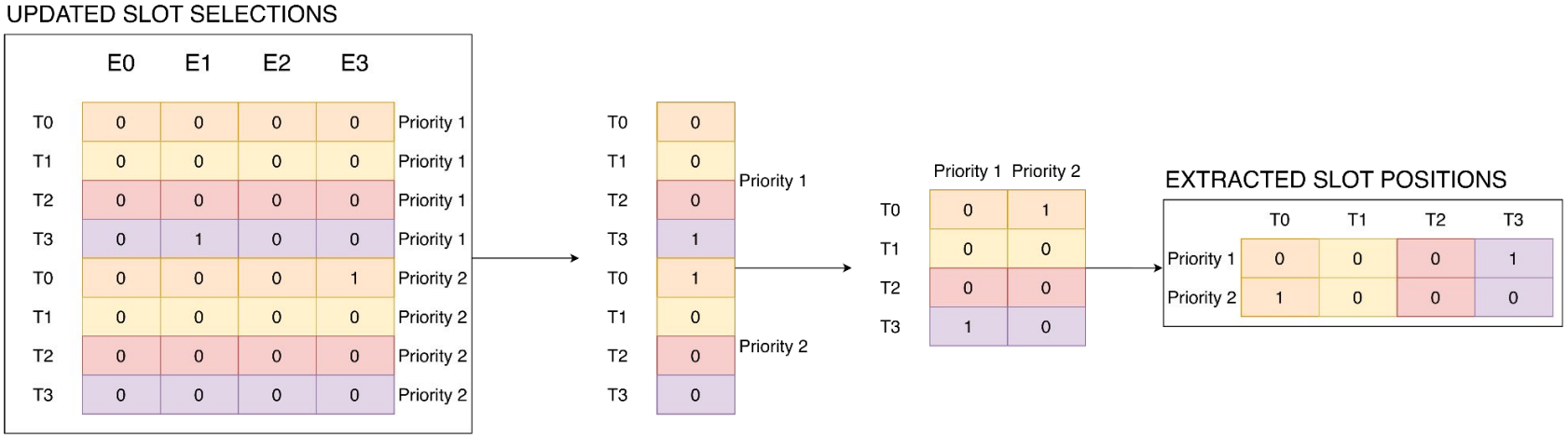


Update the Slot Selection

Extract Token Slots: *Slot positions* & *Updated selections*



Extracted Token Slots



Token	1st Choice Queue Position	2nd Choice Queue Position
Token 0	0	1
Token 1	0	0
Token 2	0	0 (dropped)
Token 3	1	0 (dropped)

Weight the token slots

What do we have till now?

UPDATED PRIORITIZED SELECTIONS

	E0	E1	E2	E3	
T0	0	1	0	0	Priority 1
T1	1	0	0	0	Priority 1
T2	0	0	0	1	Priority 1
T3	0	1	0	0	Priority 1
T0	0	0	0	1	Priority 2
T1	0	0	1	0	Priority 2
T2	0	0	0	0	Priority 2
T3	0	0	0	0	Priority 2

***Expert* to *Token* mapping**

UPDATED NORMALIZED TOKEN WEIGHTS

	E0	E1	E2	E3	
T0	0	0.5	0	0	Priority 1
T1	0.5	0	0	0	Priority 1
T2	0	0	0	0.5	Priority 1
T3	0	0.5	0	0	Priority 1
T0	0	0	0	0.4	Priority 2
T1	0	0	0.4	0	Priority 2
T2	0	0	0	0	Priority 2
T3	0	0	0	0	Priority 2

***Expert* to *Token* weights**

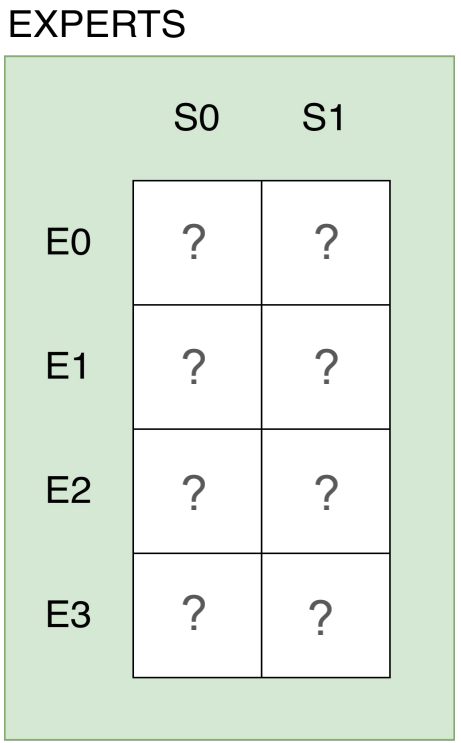
EXTRACTED SLOT POSITIONS

	T0	T1	T2	T3
Priority 1	0	0	0	1
Priority 2	1	0	0	0

***Token* slots based on Priority**

What do we need to know?

Where within each expert's capacity each token sits.



One hot slot positions

EXTRACTED SLOT POSITIONS

	T0	T1	T2	T3
Priority 1	0	0	0	1
Priority 2	1	0	0	0

SLOT ONE HOT

	Slot 0	Slot 1	
T0	1	0	Priority 1
T1	1	0	Priority 1
T2	1	0	Priority 1
T3	0	1	Priority 1
T0	0	1	Priority 2
T1	1	0	Priority 2
T2	1	0	Priority 2
T3	1	0	Priority 2

What do we have till now?

UPDATED PRIORITIZED SELECTIONS

	E0	E1	E2	E3	
T0	0	1	0	0	Priority 1
T1	1	0	0	0	Priority 1
T2	0	0	0	1	Priority 1
T3	0	1	0	0	Priority 1
T0	0	0	0	1	Priority 2
T1	0	0	1	0	Priority 2
T2	0	0	0	0	Priority 2
T3	0	0	0	0	Priority 2

***Expert* to *Token* mapping**

UPDATED NORMALIZED TOKEN WEIGHTS

	E0	E1	E2	E3	
T0	0	0.5	0	0	Priority 1
T1	0.5	0	0	0	Priority 1
T2	0	0	0	0.5	Priority 1
T3	0	0.5	0	0	Priority 1
T0	0	0	0	0.4	Priority 2
T1	0	0	0.4	0	Priority 2
T2	0	0	0	0	Priority 2
T3	0	0	0	0	Priority 2

***Expert* to *Token* weights**

EXTRACTED SLOT POSITIONS

	T0	T1	T2	T3
Priority 1	0	0	0	1
Priority 2	1	0	0	0

***Token* slots based on Priority**

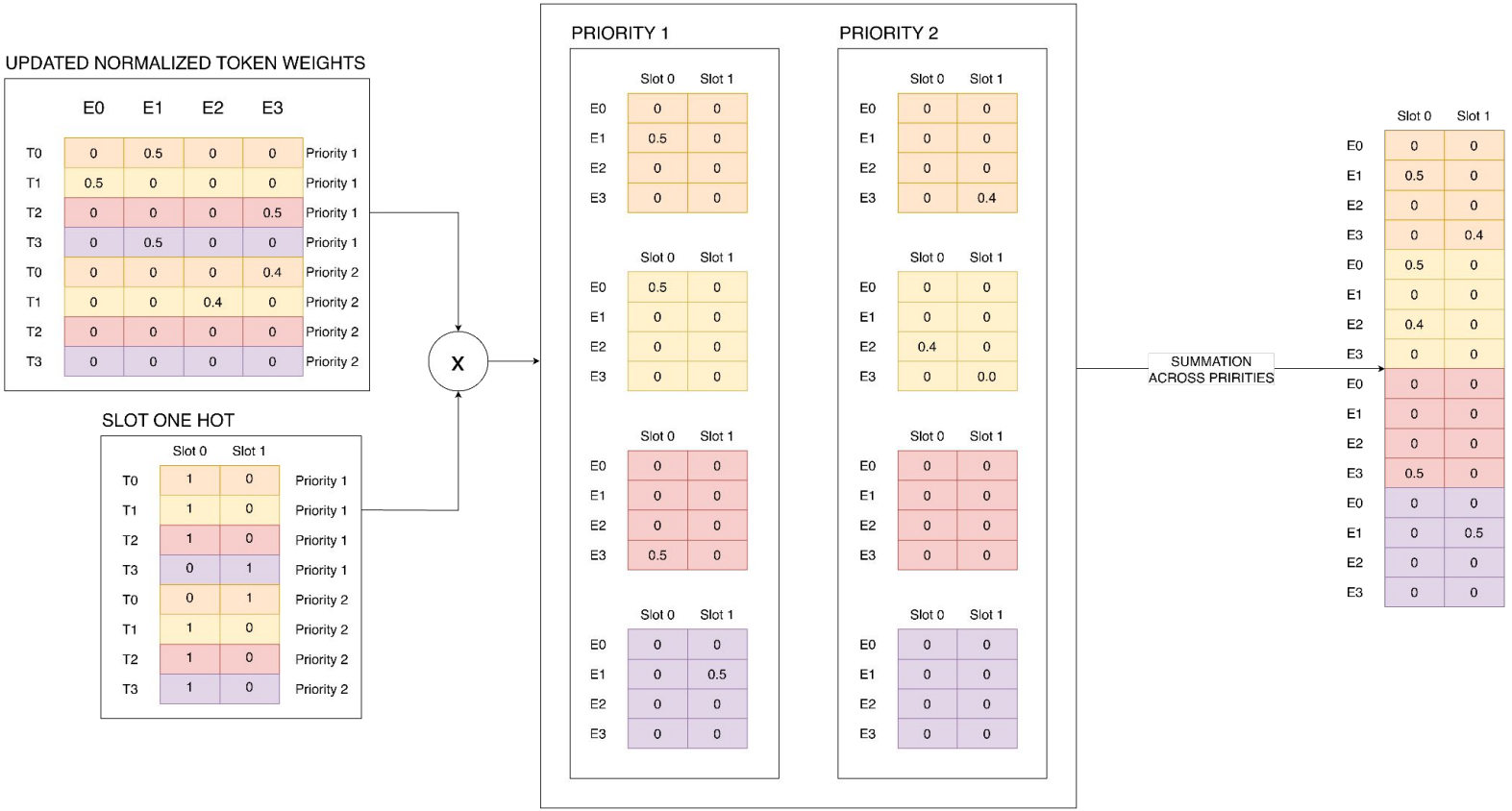
SLOT ONE HOT

	Slot 0	Slot 1	
T0	1	0	Priority 1
T1	1	0	Priority 1
T2	1	0	Priority 1
T3	0	1	Priority 1
T0	0	1	Priority 2
T1	1	0	Priority 2
T2	1	0	Priority 2
T3	1	0	Priority 2

***Token* to *Slot* mapping**

(based on priority)

Token weights in each slot position



Token weights in each slot position

	Slot 0	Slot 1
E0	0	0
E1	0.5	0
E2	0	0
E3	0	0.4
E0	0.5	0
E1	0	0
E2	0.4	0
E3	0	0
E0	0	0
E1	0	0
E2	0	0
E3	0.5	0
E0	0	0
E1	0	0.5
E2	0	0
E3	0	0

	Slot 0	Slot 1
E0	0.5 x Token 1	
E1	0.5 x Token 0	0.5 x Token 3
E2	0.4 x Token 1	
E3	0.5 x Token 2	0.4 x Token 0

Bring it home!

Place tokens in each Expert's Slot