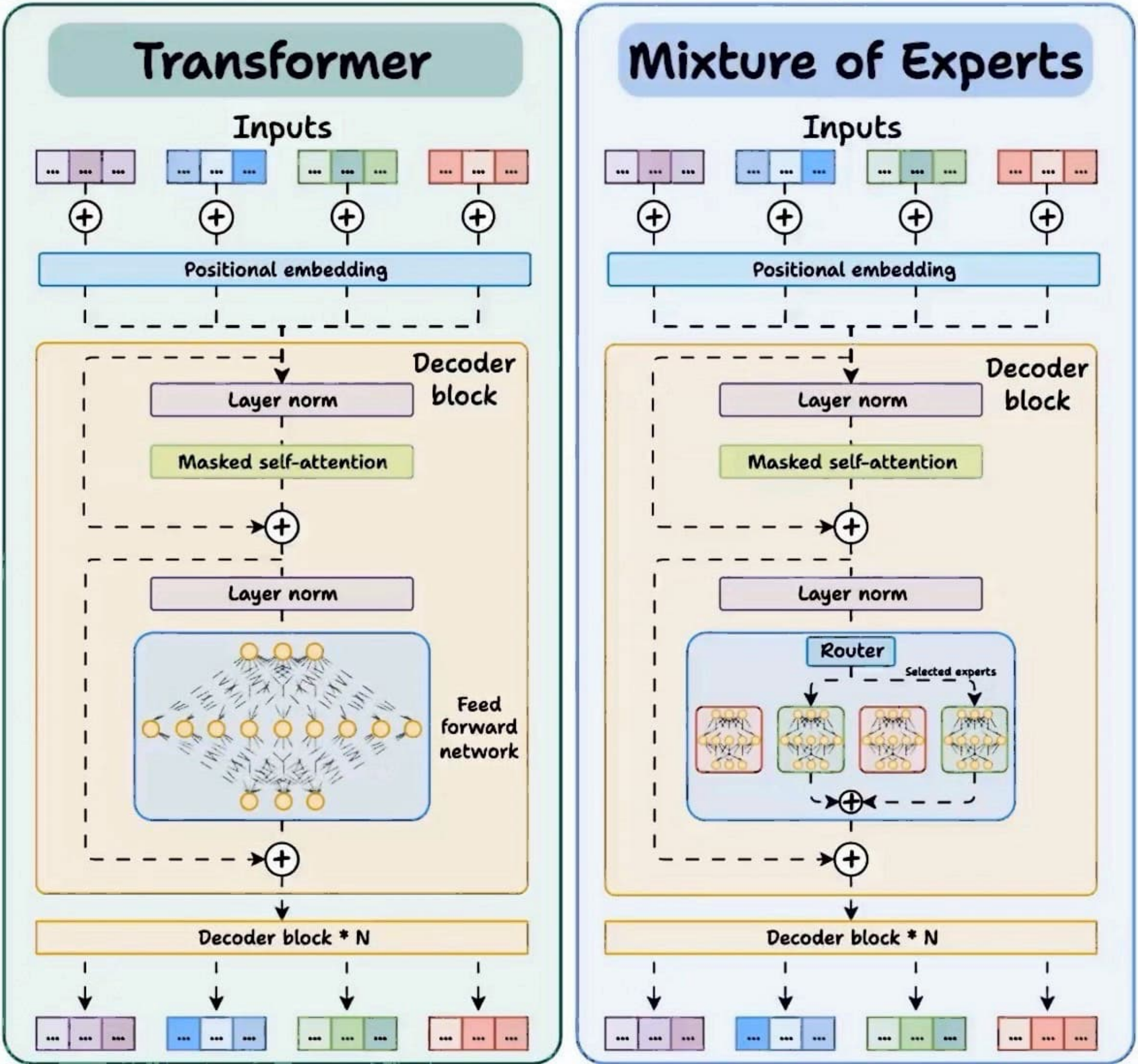
Transformer vs. Mixture of Experts

join.DailyDoseofDS.com

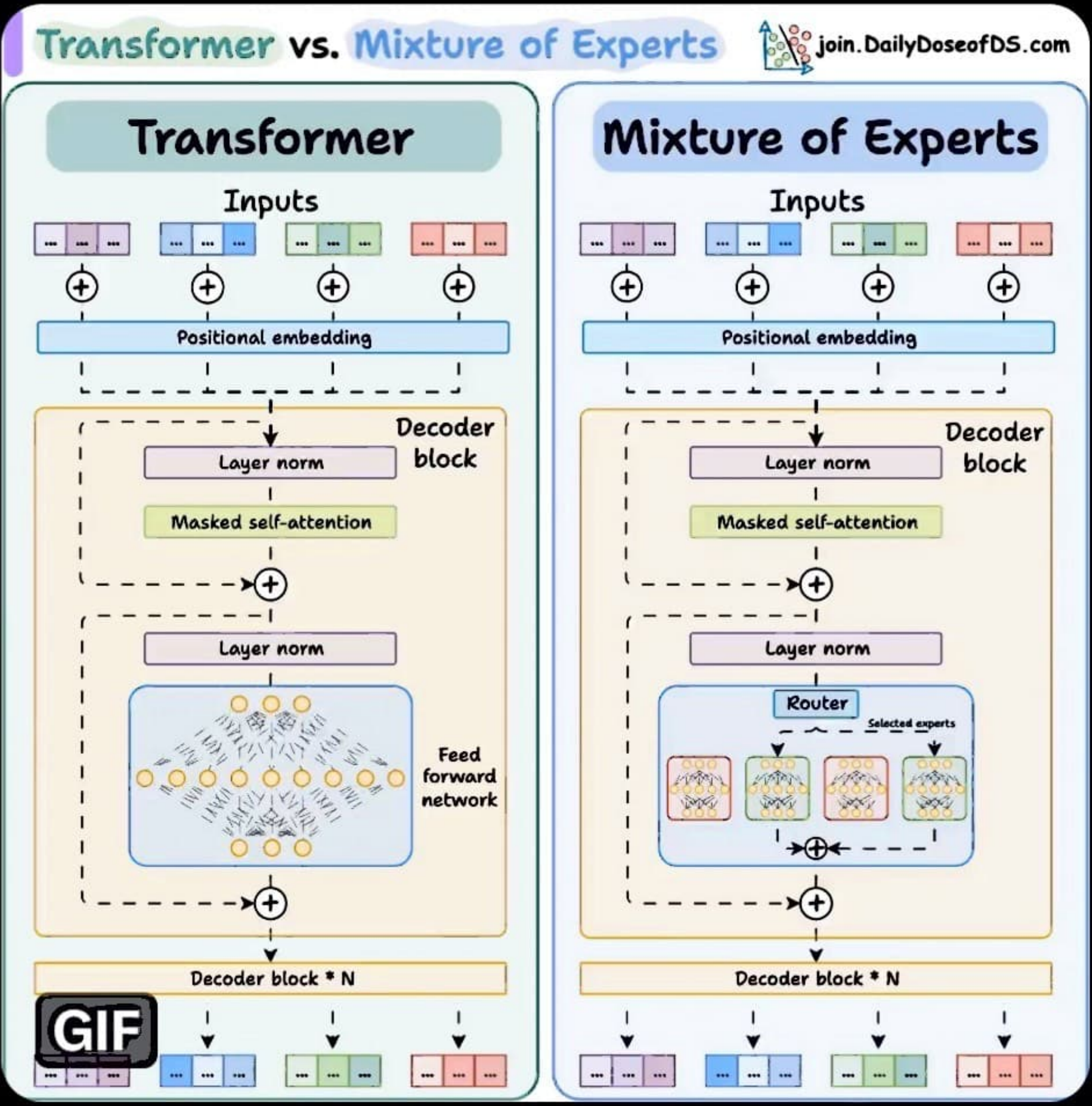


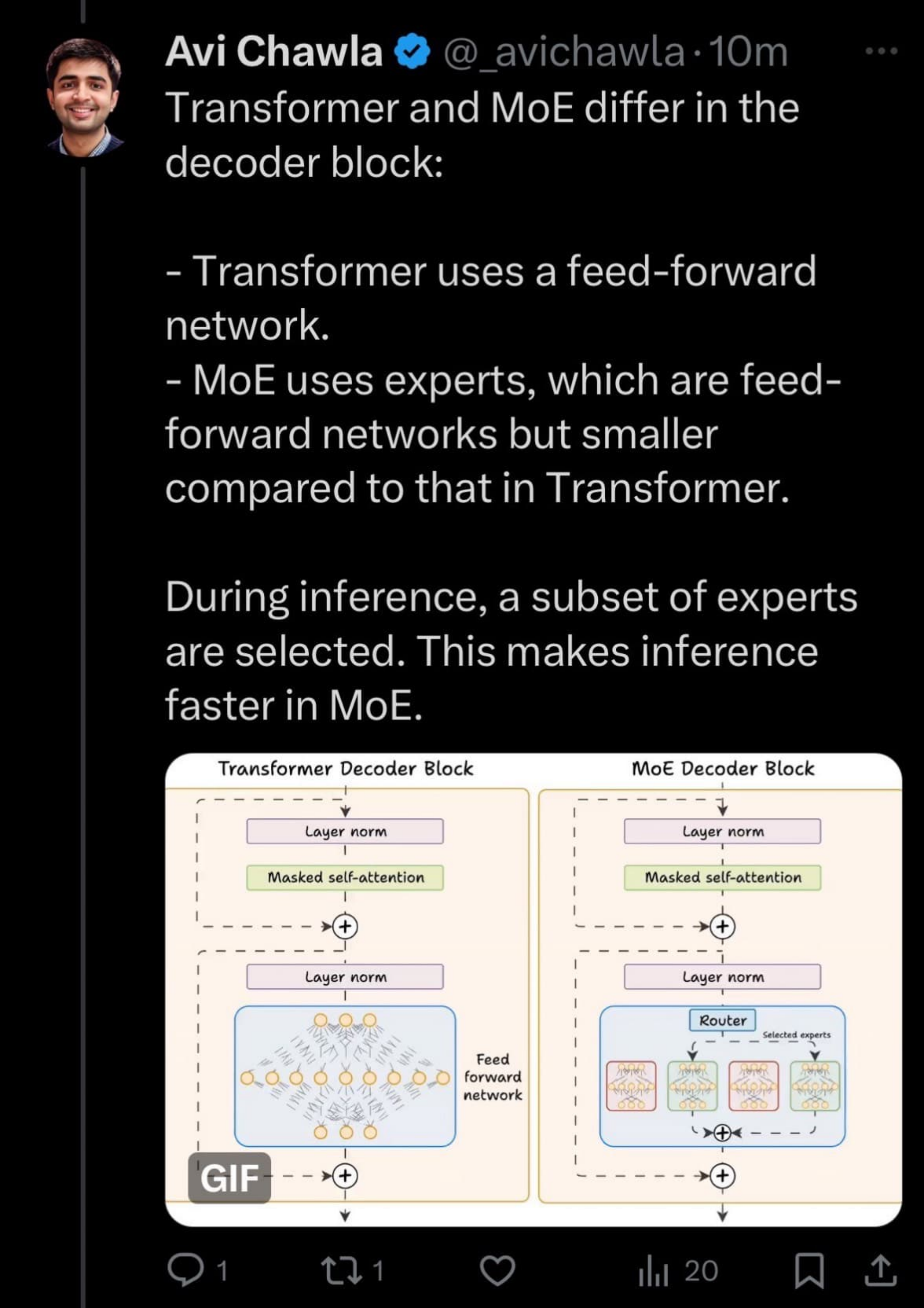
Avi Chawla @_avichawla · 10m

Mixture of Experts (MoE) is a popular architecture that uses different "experts" to improve Transformer models.

The visual below explains how they differ from Transformers.

Let's dive in to learn more about MoE!



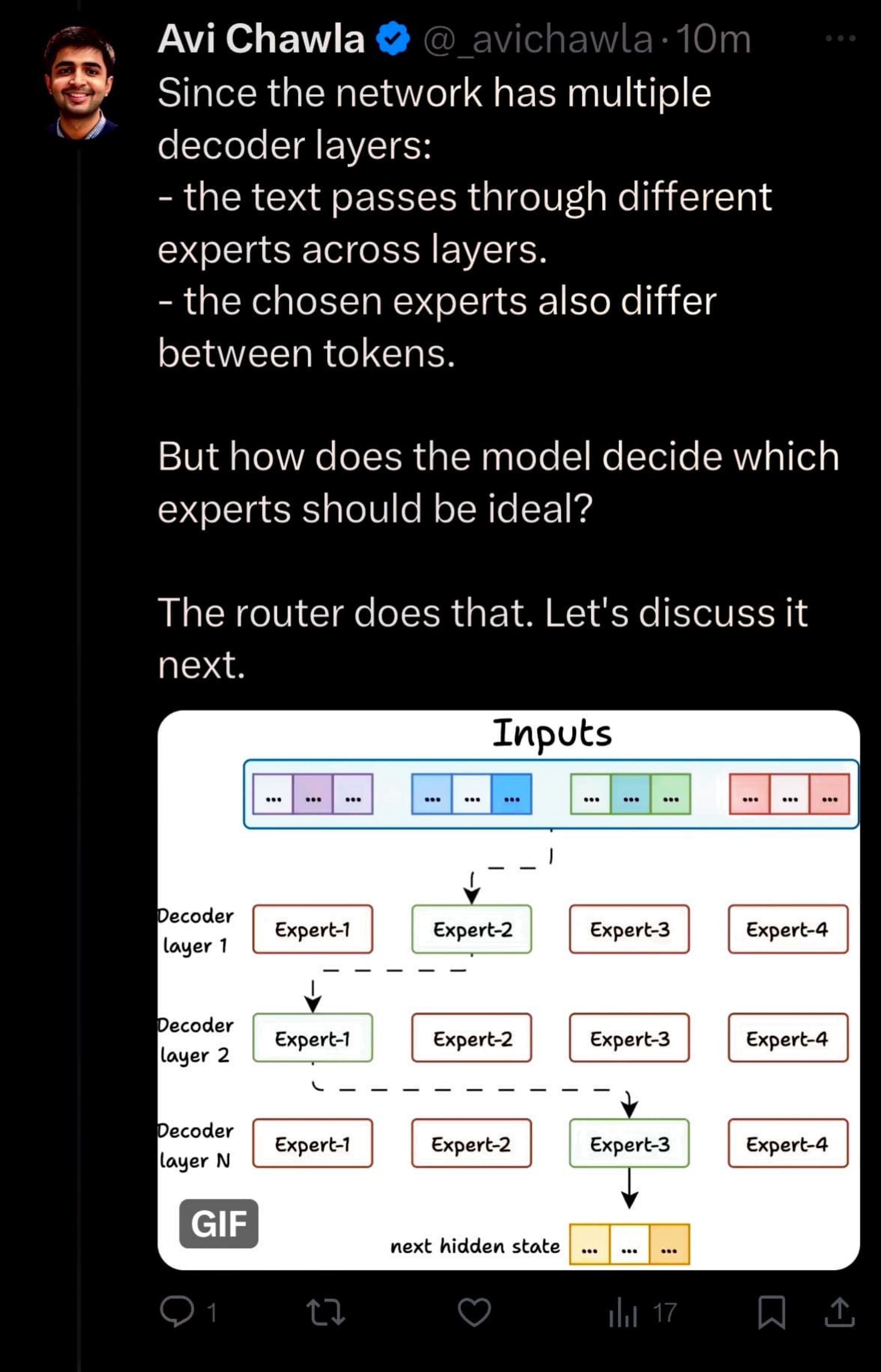


Avi Chawla @_avichawla · 10m

Transformer and MoE differ in the decoder block:

- Transformer uses a feed-forward network.
- MoE uses experts, which are feed-forward networks but smaller compared to that in Transformer.

During inference, a subset of experts are selected. This makes inference faster in MoE.

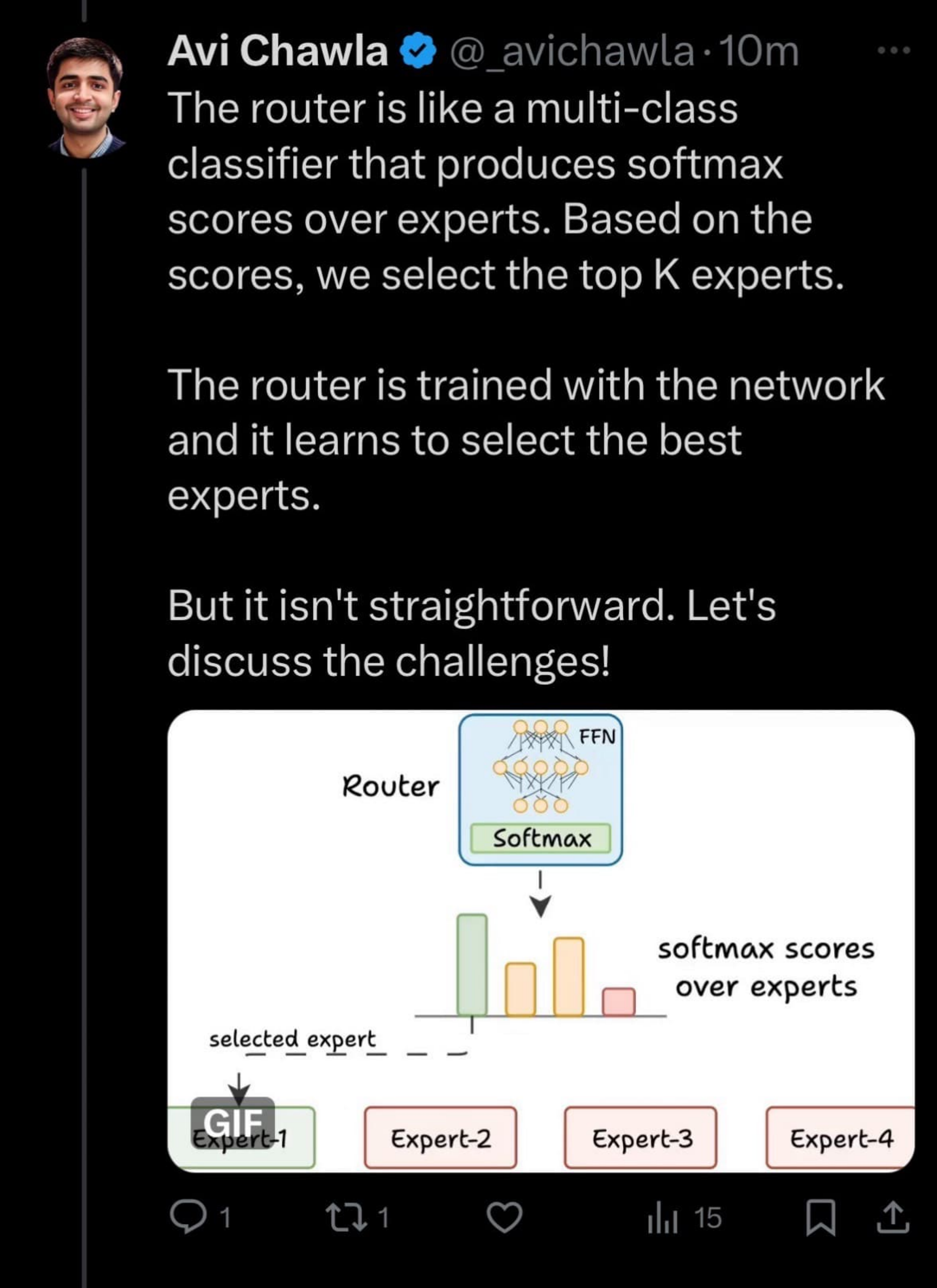


Avi Chawla @_avichawla · 10m

Since the network has multiple decoder layers:
- the text passes through different experts across layers.
- the chosen experts also differ between tokens.

But how does the model decide which experts should be ideal?

The router does that. Let's discuss it next.

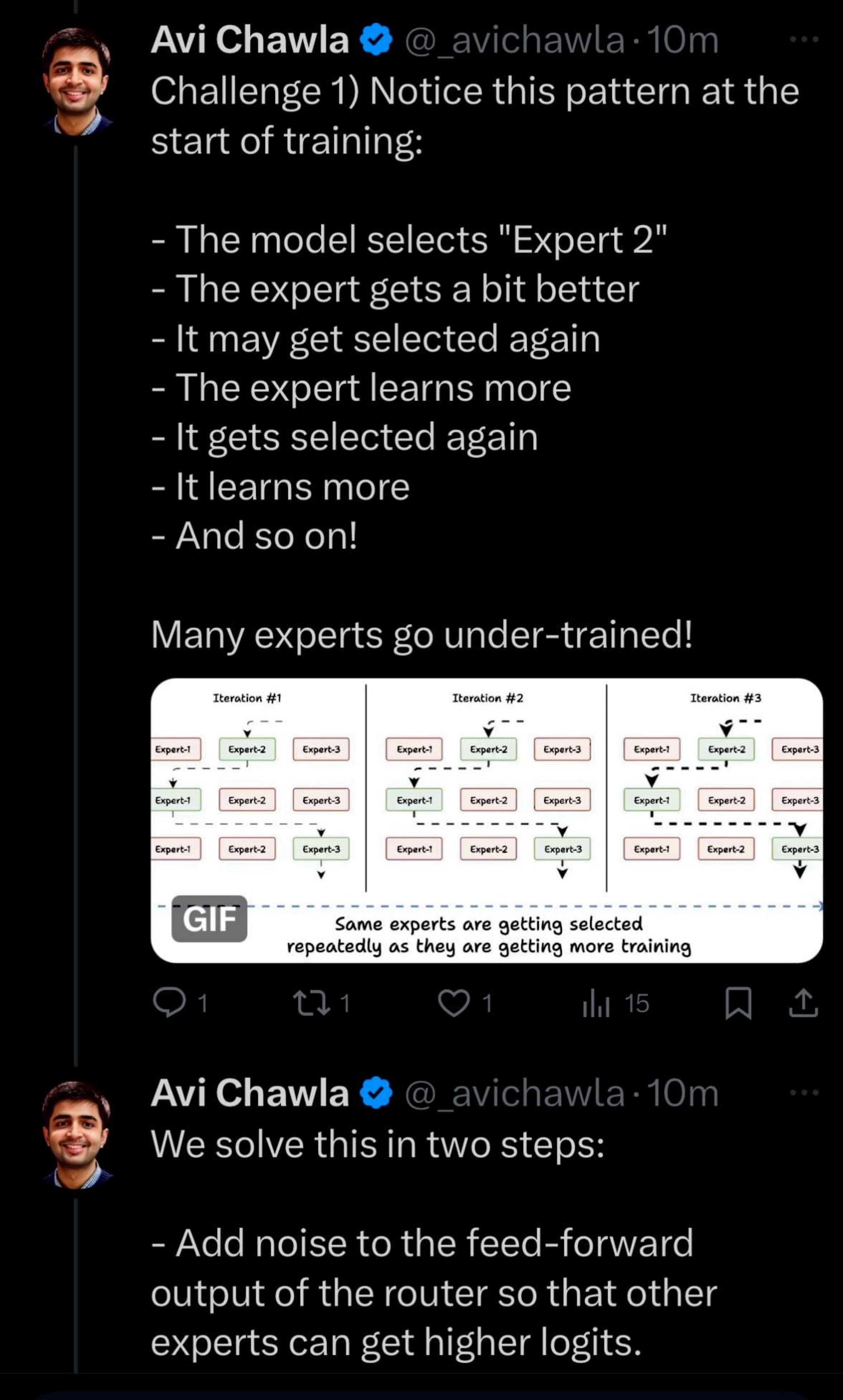


Avi Chawla @_avichawla · 10m

The router is like a multi-class classifier that produces softmax scores over experts. Based on the scores, we select the top K experts.

The router is trained with the network and it learns to select the best experts.

But it isn't straightforward. Let's discuss the challenges!



Avi Chawla @_avichawla · 10m

Challenge 1) Notice this pattern at the start of training:

- The model selects "Expert 2"
- The expert gets a bit better
- It may get selected again
- The expert learns more
- It gets selected again
- It learns more
- And so on!

Many experts go under-trained!

Same experts are getting selected repeatedly as they are getting more training

Avi Chawla @_avichawla · 10m

We solve this in two steps:

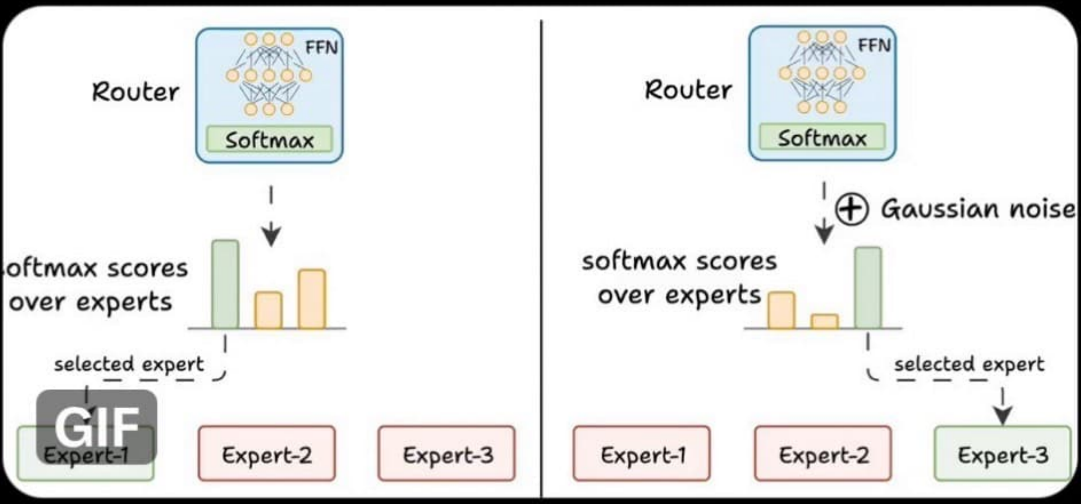
- Add noise to the feed-forward output of the router so that other experts can get higher logits.

Avi Chawla @_avichawla · 10m

We solve this in two steps:

- Add noise to the feed-forward output of the router so that other experts can get higher logits.
- Set all but top K logits to -infinity. After softmax, these scores become zero.

This way, other experts also get the opportunity to train.

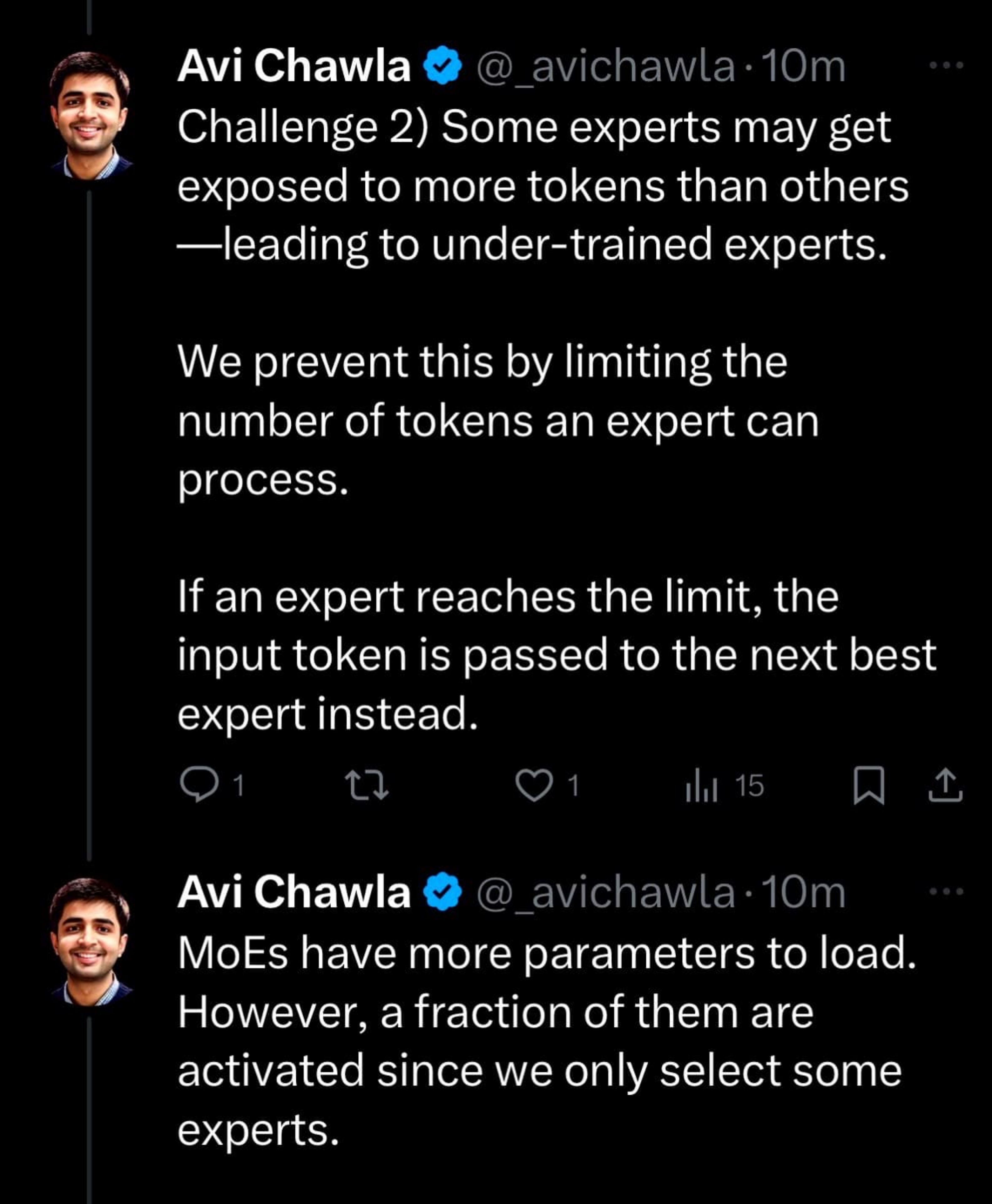


1 1 1 18

Avi Chawla @_avichawla · 10m

Challenge 2) Some experts may get exposed to more tokens than others—leading to under-trained experts.

We prevent this by limiting the number of tokens an expert can



Avi Chawla @_avichawla · 10m

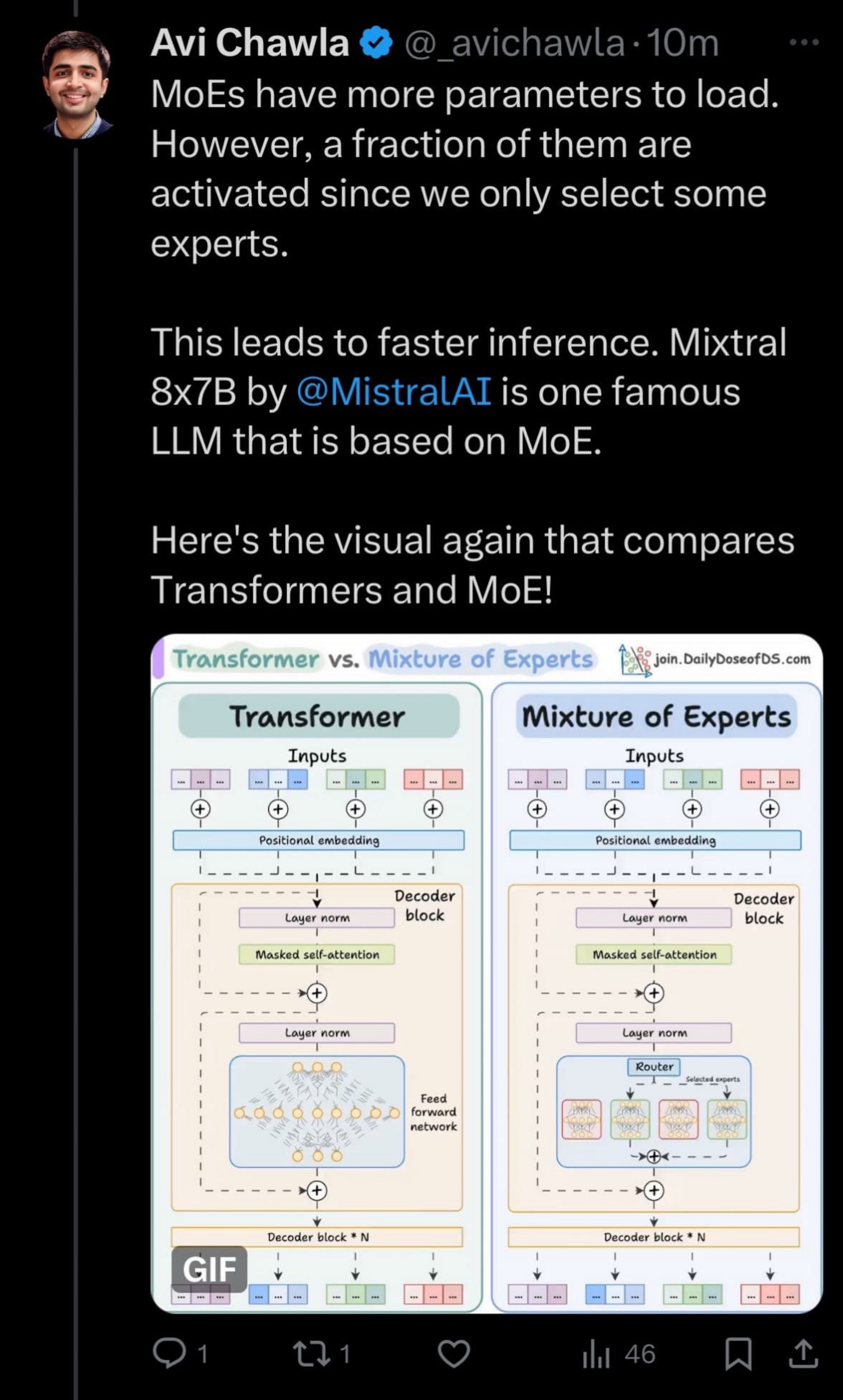
Challenge 2) Some experts may get exposed to more tokens than others —leading to under-trained experts.

We prevent this by limiting the number of tokens an expert can process.

If an expert reaches the limit, the input token is passed to the next best expert instead.

Avi Chawla @_avichawla · 10m

MoEs have more parameters to load. However, a fraction of them are activated since we only select some experts.



Avi Chawla @_avichawla · 10m

MoEs have more parameters to load. However, a fraction of them are activated since we only select some experts.

This leads to faster inference. Mixtral 8x7B by @MistralAI is one famous LLM that is based on MoE.

Here's the visual again that compares Transformers and MoE!

1 1 46