

Scaling Laws for Upcycling Mixture-of-Experts Language Models

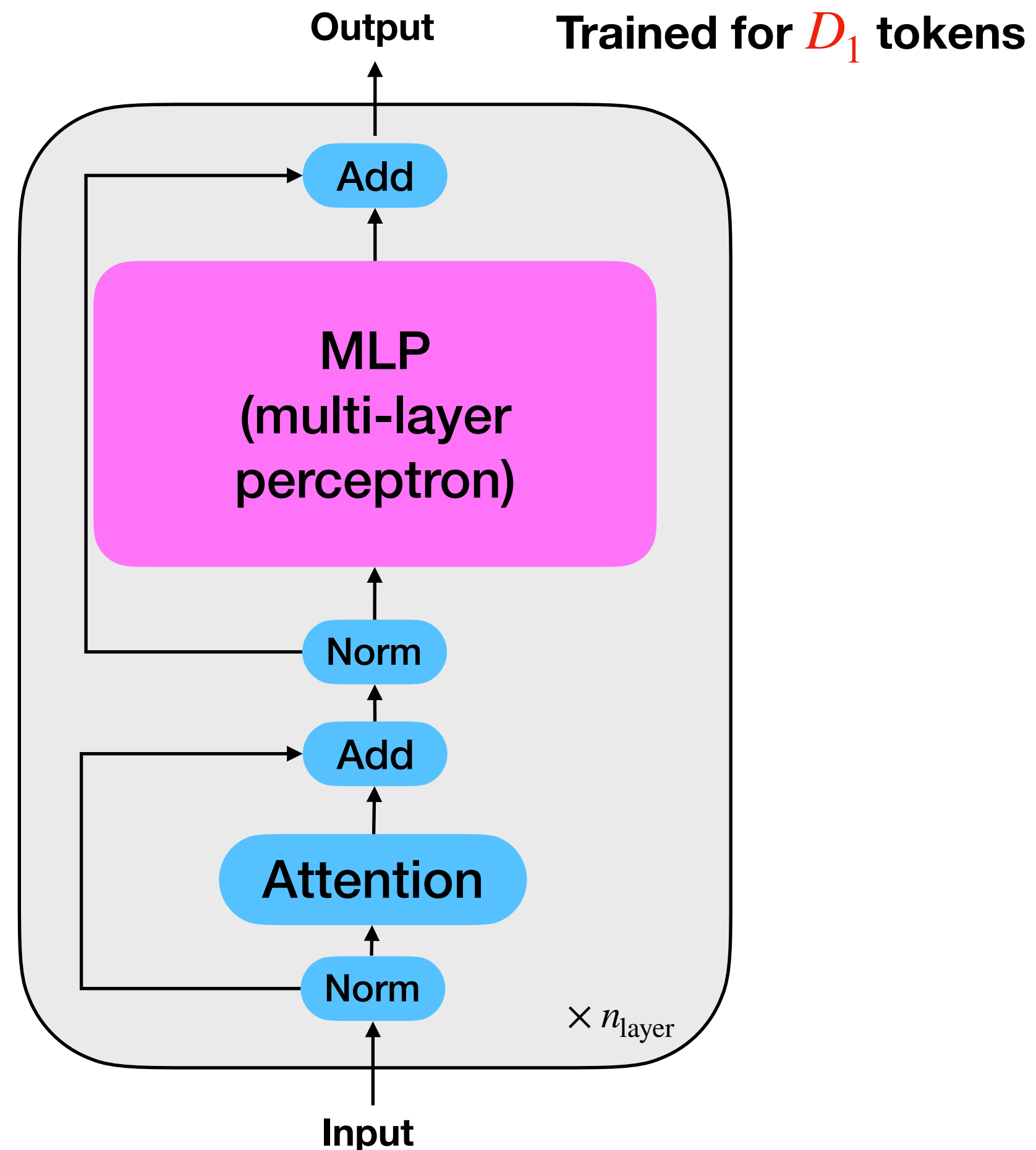
TLDR: Overtrained dense (language) transformers are *harder* to upcycle to Mixture-of-Experts (MoE) models

Seng Pei Liew, Takuya Kato, Sho Takase

What is upcycling?

[2212.05055] Reuse pretrained dense models to train MoE efficiently

Dense model (Llama etc.) of size N_1

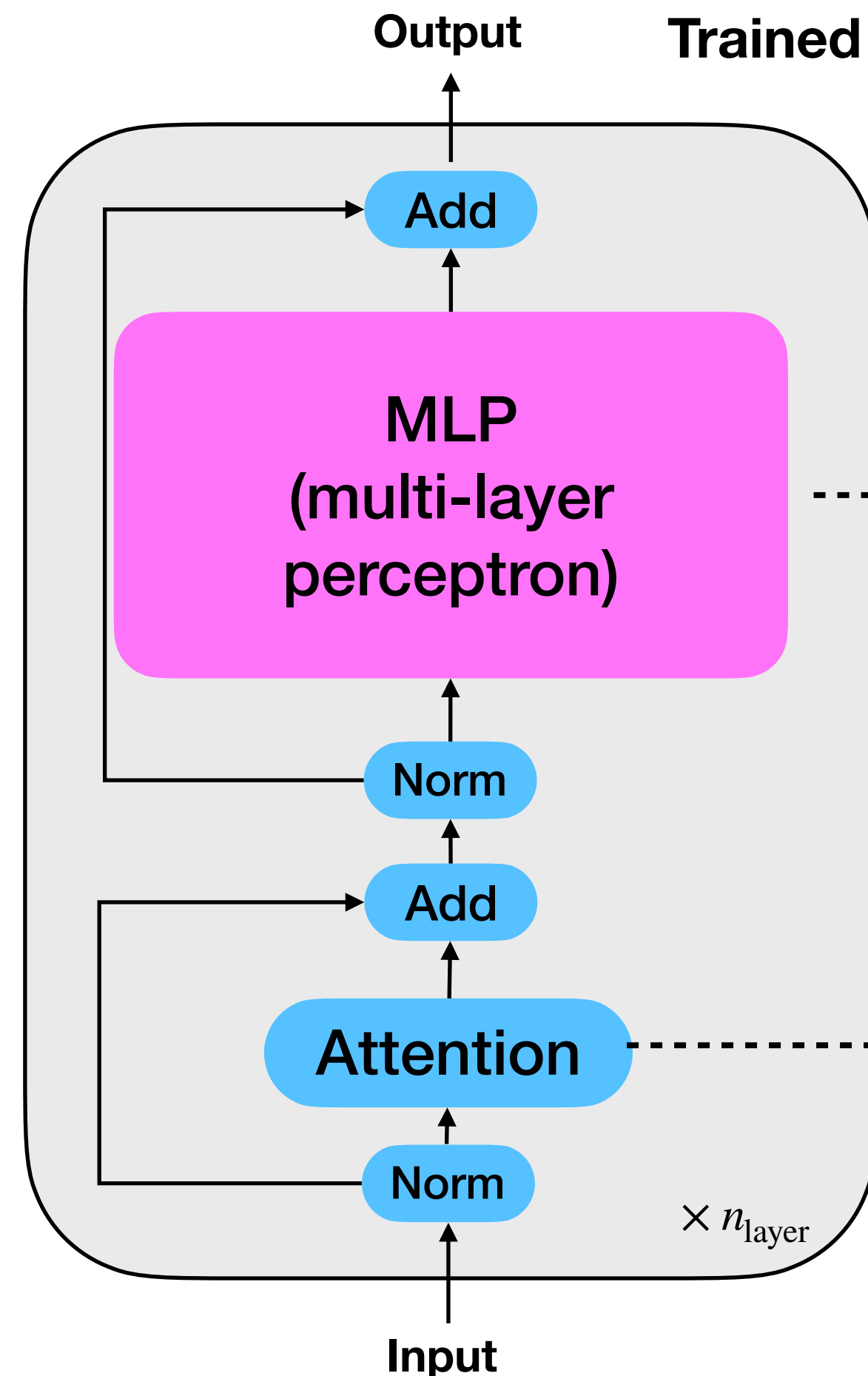


What is upcycling?

[2212.05055] Reuse pretrained dense models to train MoE efficiently

Dense model (Llama etc.) of size N_1

Trained for D_1 tokens

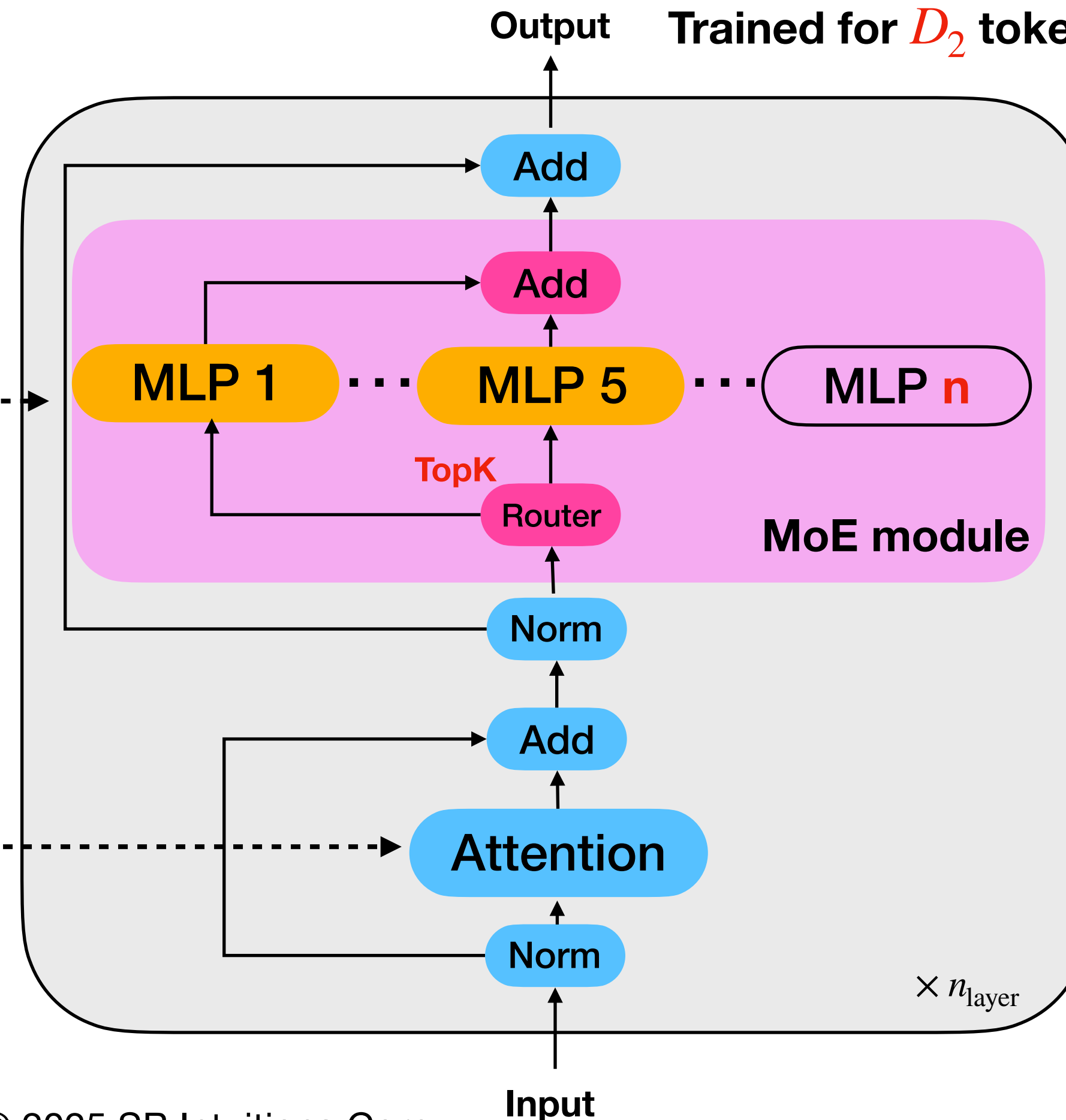


Duplicate

Copy

Upcycled MoE (Mixtral etc.)


Trained for D_2 tokens



Scaling law for dataset sizes fixing model configuration

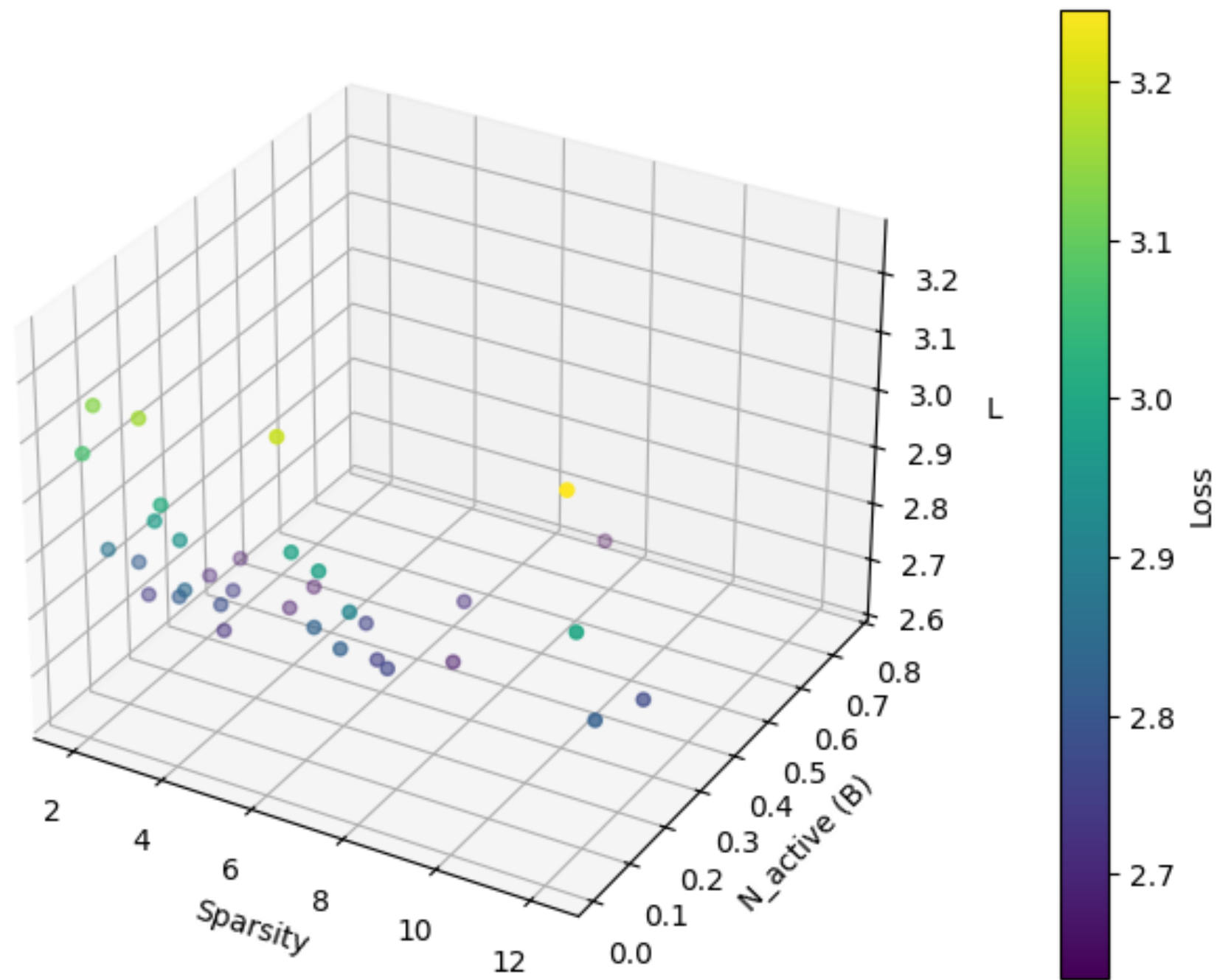
Dense model (Llama etc.) of size N_1 Trained for D_1 tokens

Upcycled MoE (Mixtral etc.) Trained for D_2 tokens

$$L = A D_1^{-\alpha_1} D_2^{-\alpha_2 + \alpha_3 \log D_1} + E \quad (\alpha_2 > \alpha_1 > \alpha_3)$$


Interaction term: the more overtrained the dense model is, the harder the MoE can be upcycled/trained

Scaling law for model configuration fixing dataset sizes

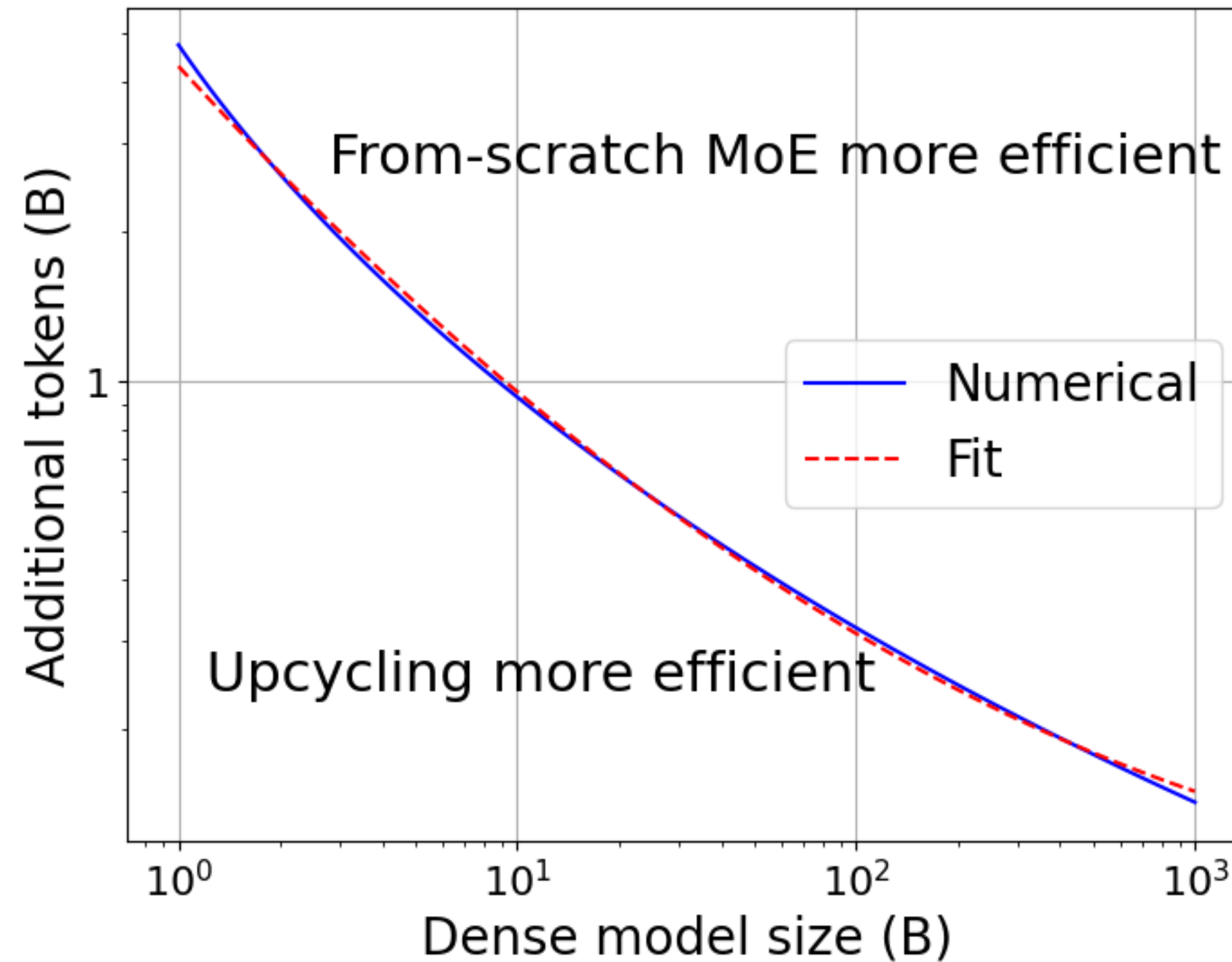


$$L = AP^{-\beta_1} + FN_2^{-\beta_2} + E$$

The sparser (MoE) and the larger (base model) the model is, the better

Joint scaling law for Mixtral-like MoE

$$L = AD_1^{-\alpha_1}D_2^{-\alpha_2+\alpha_3\log D_1} + BN_1^{-\beta_2} + E$$



Joint scaling law provides guidance on when to upcycle/train from scratch

Check out our paper/poster for details!

Preprint:



Poster:



 **SB Intuitions**