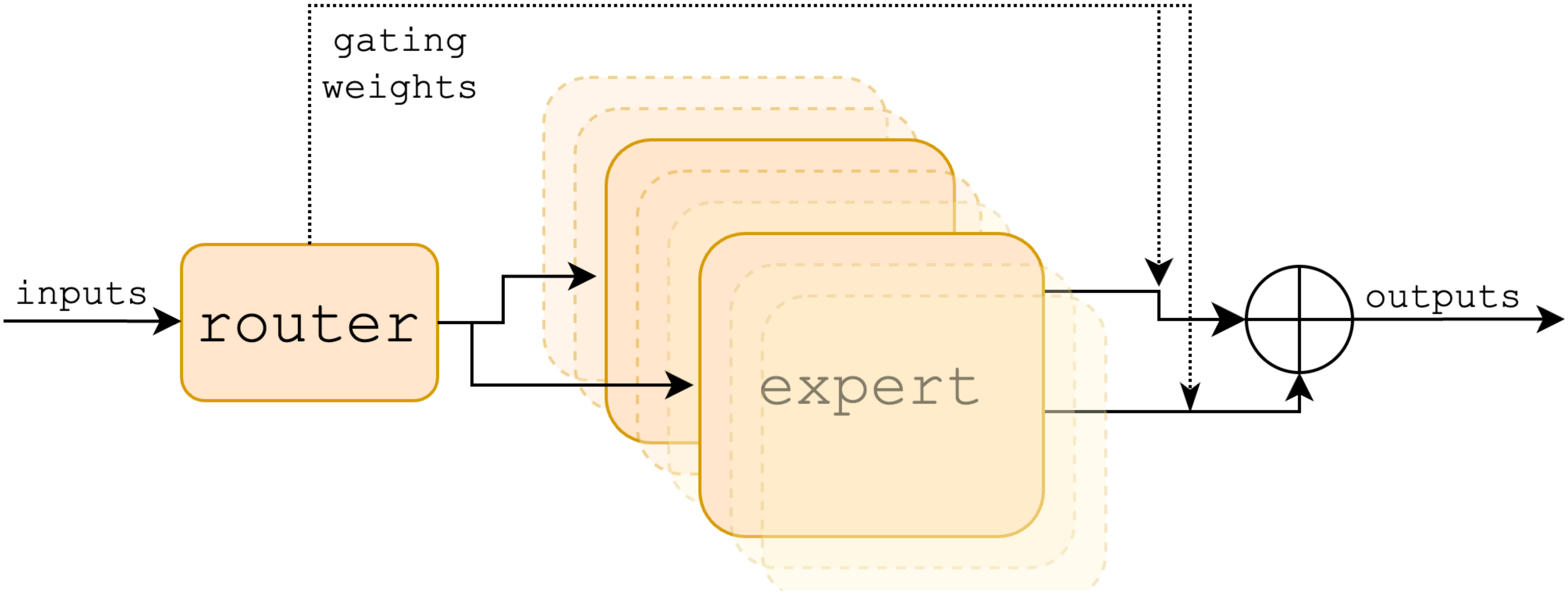
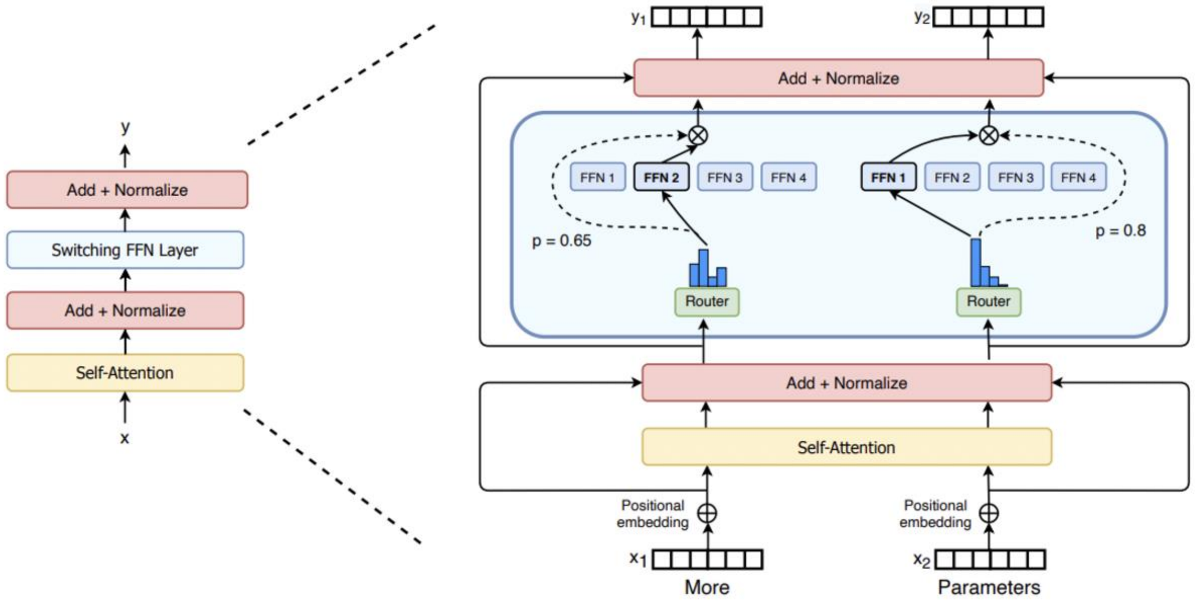


Mixtral of Experts

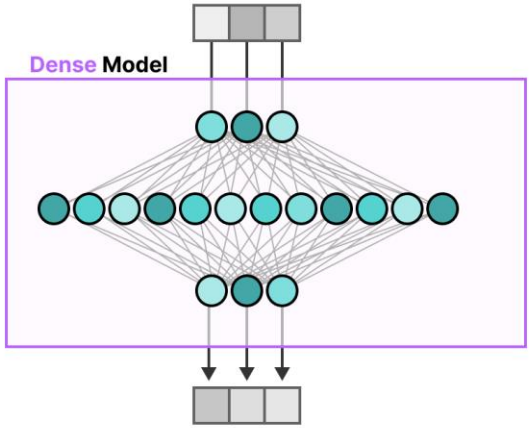
MIAN ZHONG, ZIHAO ZHAO

What is Sparse Mixture of Experts (MoE)?

Every FFN layer of the transformer model is replaced with an MoE layer.
MoE layers have a certain number of “experts”, where each expert is a neural network (FFN).
The **router** determines which tokens are sent to which expert.



$$y = \sum_{i=1}^{n} G(x)_i E_i(x)$$



What is Sparsity?

While in dense models all the parameters are used for all the inputs, sparsity allows us to only run some parts of the whole system.

Wg: learned parameter matrix

$$(\mathrm{TopK}(\ell))_i = \begin{cases} \ell_i & \text{if } \ell_i \text{ is among the top-k values} \\ -\infty & \text{otherwise.} \end{cases}$$

$$G(x) \coloneqq Softmax(TopK(x \cdot W_g))$$

Expert initialization

(1) All experts are initialized identically using a common initialization scheme

(2) Train each expert independently on its assigned data or task

(3) Sparse Upcycling Initialization:

- Start with a dense model pre-trained on a general task.
- "Split" the dense model's parameters into multiple experts, assigning a subset of the dense model's parameters to each expert.
- Fine-tune the sparse MoE model on the downstream task.

Advantages of MoE

Efficient Training: Only the active experts (e.g., Top-k) contribute to each computation, saving memory and computation time.

Lower computational cost during training and inference

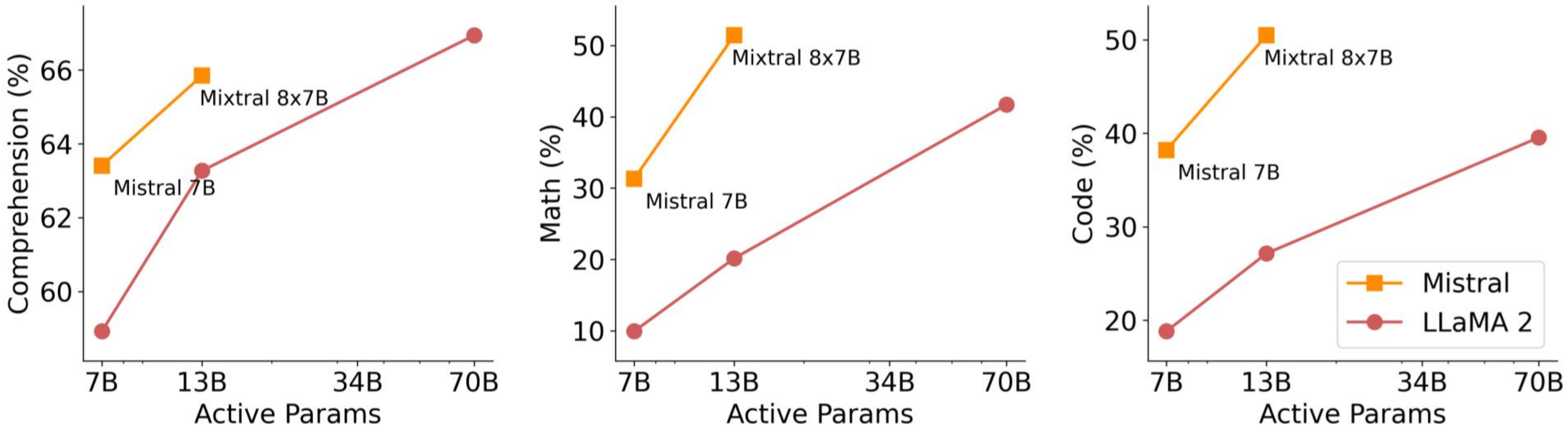
Different experts can specialize in handling specific patterns, tasks, or domains in the data. This improves the model's ability to generalize and perform well across a wide variety of inputs.

Limitations of MoE

- All parameters need to be loaded in RAM, so memory requirements are high. For example, given a MoE like Mixtral 8x7B, we'll need to have enough VRAM to hold a dense 47B parameter model.
- Some experts might be over-utilized while others remain under-utilized, leading to inefficient use of resources.

Mixtral of Experts (Mistral.AI)

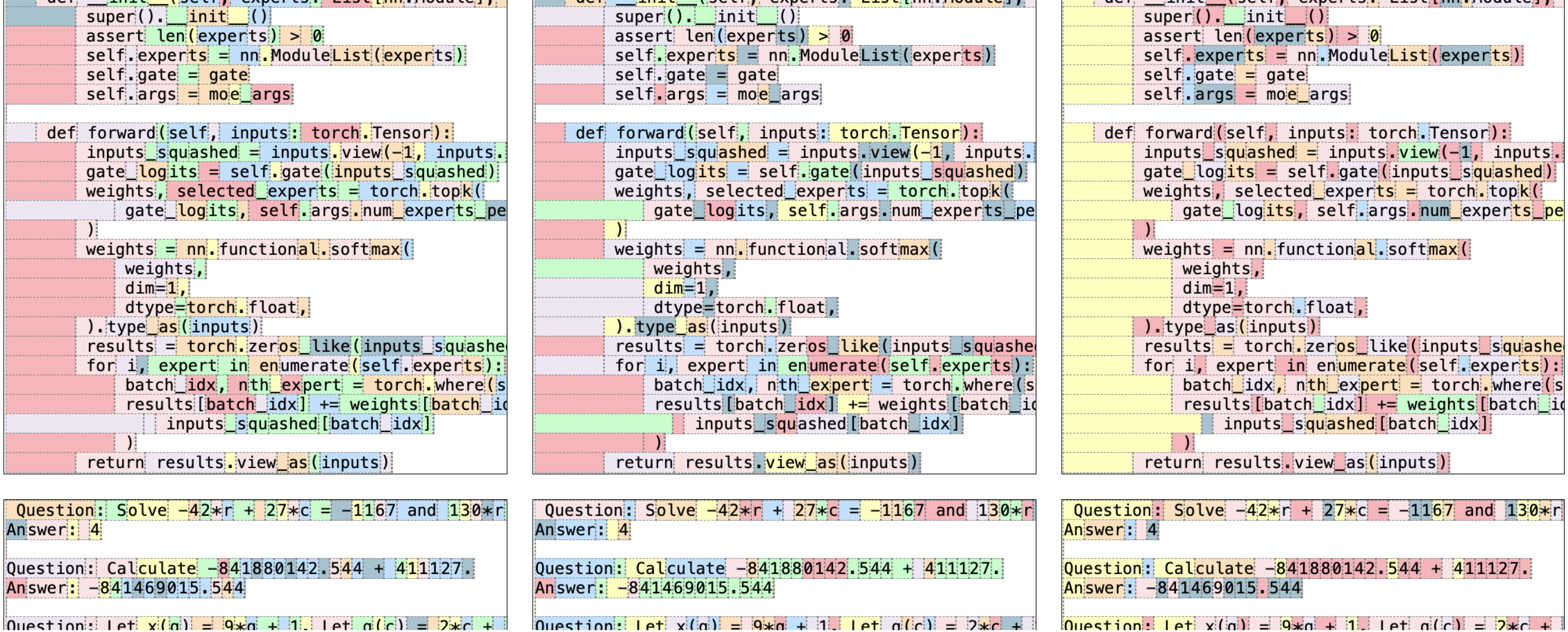
- Long context window: 32k tokens
- > Llama 2 70B with 6x faster inference
- Math/Code Generation 👍 Multilingual (upsampling pre-train) 👍 Less bias 👍



Design Choice

- More experts, same computation, more memory transfer
- Why Top-2 routing?
  - Capacity Factor: expand/contract "tokens per expert"
  - Likely the best hardware-software optimization for Mixtral

Algorithm	Train CF	Eval CF	Neg. Log Perp. (↑)
Dense-L	—	—	-1.474
Dense-XL	—	—	-1.384
Top-1	0.75	0.75	-1.428
Top-1	0.75	2.0	-1.404
Top-2	0.75	0.75	-1.424
Top-2	0.75	2.0	-1.402
Top-1	1.0	1.0	-1.397
Top-1	1.0	2.0	-1.384
Top-2	1.0	1.0	-1.392
Top-2	1.0	2.0	-1.378
Top-1	1.25	1.25	-1.378
Top-1	1.25	2.0	-1.373
Top-2	1.25	1.25	-1.375
Top-2	1.25	2.0	-1.369
Top-2	2.0	2.0	-1.360
Top-2	2.0	3.0	-1.359
Top-3	2.0	2.0	-1.360
Top-3	2.0	3.0	-1.356



What are experts specialized in?

😮 Syntax, not domains.

More on experts specialization

- Encoder-decoder model ST-MoE pretrained on C4
- More obvious at encoder
  - An expert on mask tokens
- Far less noticeable in the decoder
- Multilingual experts are not by languages
  - Routers pass indiscriminately -> all experts are encouraged to handle all languages

Repeated Experts

	First choice			First or second choice		
	Layer 0	Layer 15	Layer 31	Layer 0	Layer 15	Layer 31
ArXiv	14.0%	27.9%	22.7%	46.5%	62.3%	52.9%
DM Mathematics	14.1%	28.4%	19.7%	44.9%	67.0%	44.5%
Github	14.9%	28.1%	19.7%	49.9%	66.9%	49.2%
Gutenberg	13.9%	26.1%	26.3%	49.5%	63.1%	52.2%
PhilPapers	13.6%	25.3%	22.1%	46.9%	61.9%	51.3%
PubMed Abstracts	14.2%	24.6%	22.0%	48.6%	61.6%	51.8%
StackExchange	13.6%	27.2%	23.6%	48.2%	64.6%	53.6%
Wikipedia (en)	14.4%	23.6%	25.3%	49.8%	62.1%	51.8%

- The proportion of two consecutive tokens get the same expert

Repeated Experts