Ensemble Models
[Mixture of Experts]

The Evolution of MoE

The **MoE concept isn't new**. It dates back to 1991, with significant milestones along the way:

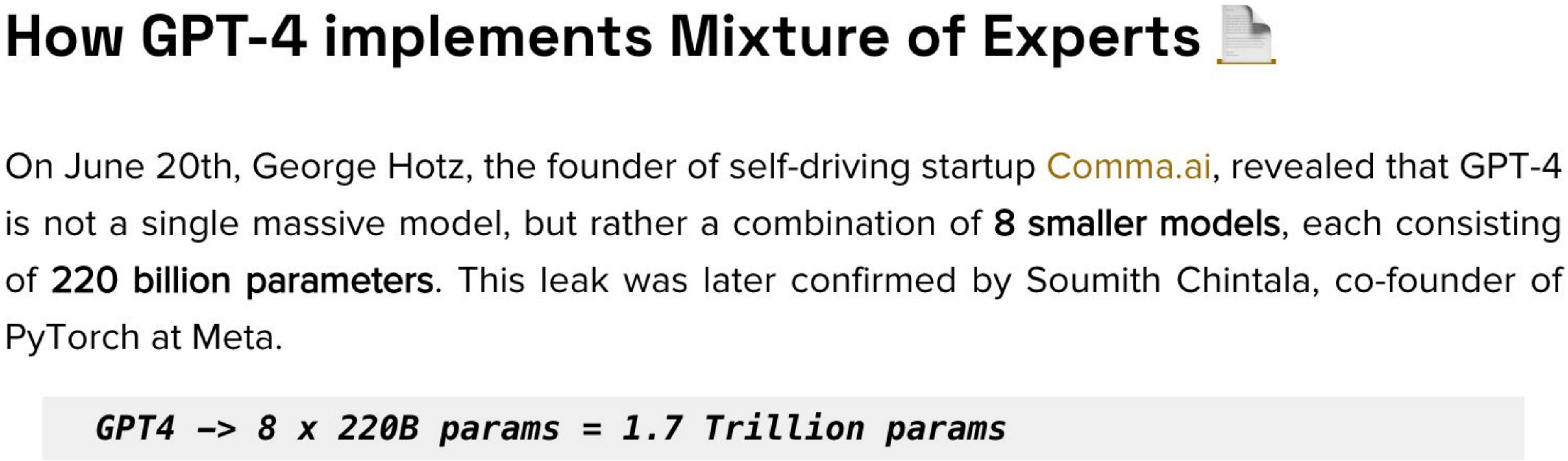
- **1991**: 📄 Geoffrey Hinton proposed Mixtures of Local Experts for the first time
- **2014**: 📄 MoE was first applied to deep learning.
- **2017**: 📄 Geoffrey Hinton proposed using MoE on large-scale models.
- **2020**: 📄 Google's GShard experimented with MoE in giant transformers.
- **2022**: 📄 Google's Switch Transformers addressed some of MoE's training and fine-tuning issues.

In the context of LLMs, the concept of 'expertise' takes a unique form. Each model, or 'expert,' naturally develops a proficiency in different topics as it undergoes the training process.

In this setup, the role of a 'coordinator,' which in a human context might be a person overseeing a team, is played by a Gating Network.

This network has the crucial task of directing inputs to the appropriate models based on the topic at hand.

Over time, the **Gating Network** improves its understanding of each model's strengths and fine-tunes its routing decisions accordingly.



How GPT-4 implements Mixture of Experts 📄

On June 20th, George Hotz, the founder of self-driving startup Comma.ai, revealed that GPT-4 is not a single massive model, but rather a combination of **8 smaller models**, each consisting of **220 billion parameters**. This leak was later confirmed by Soumith Chintala, co-founder of PyTorch at Meta.

```
GPT4 -> 8 x 220B params = 1.7 Trillion params
```

For context, **GPT-3.5 has around 175B parameters**.

However, just like we will cover in Mixtral, the calculation of the total number of parameters when using MoE is not so direct since only FFN (feed-forward network) layers are replicated between each expert, while the other layers can be shared by all.

This may significantly decrease the
total number of parameters of GPT-4.

Regardless the total number should be
somewhere between **1.2-1.7 Trillion parameters**.

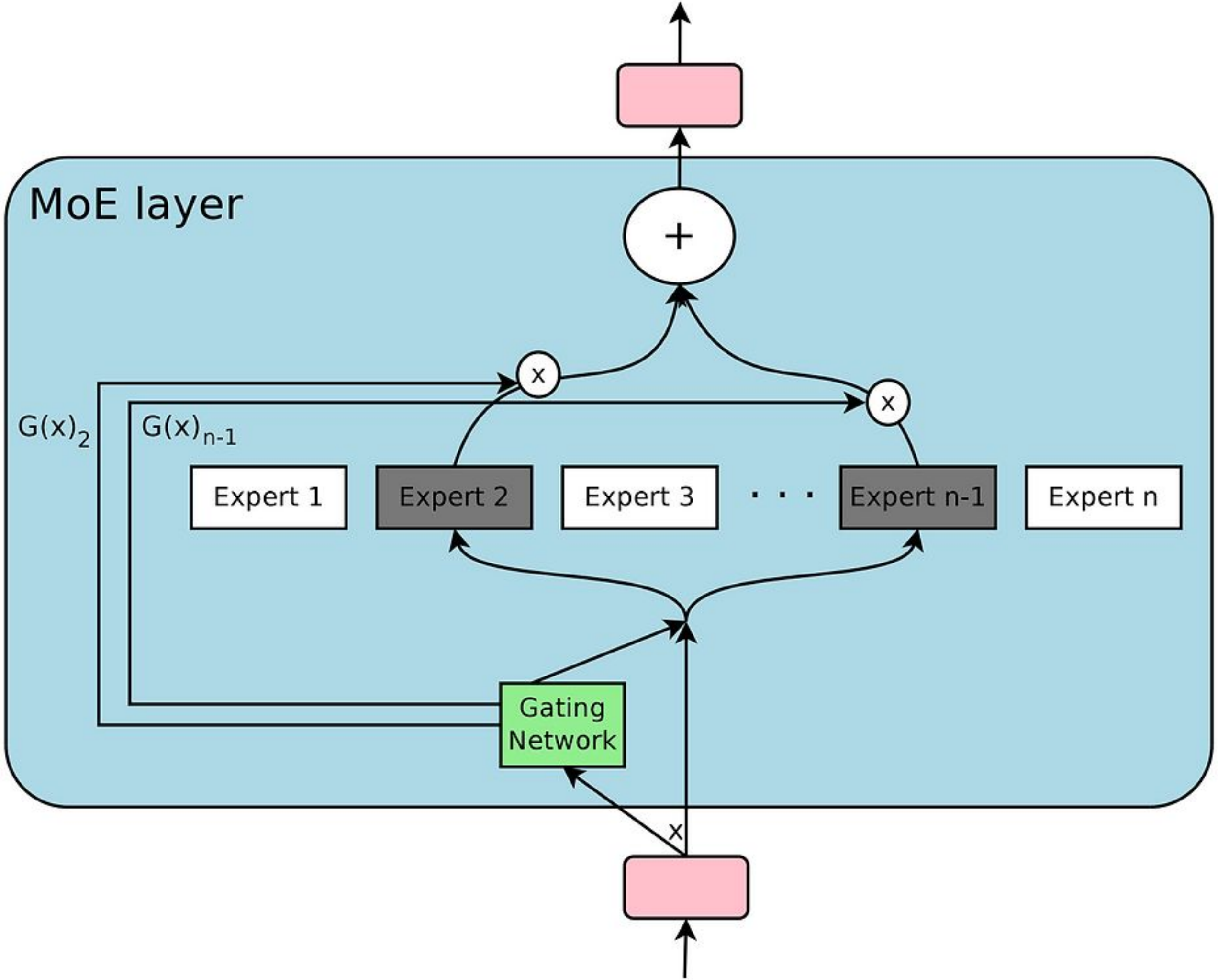
Mistral 8x7B aka Mixtral explained 📄

Mixtral is **outperforming many large models** while being efficient in inference. It employs a routing layer that decides which expert or combination of experts to use for each task, optimizing resource usage. It only has **46.7B parameters**, but only uses about **12.9B per token**.

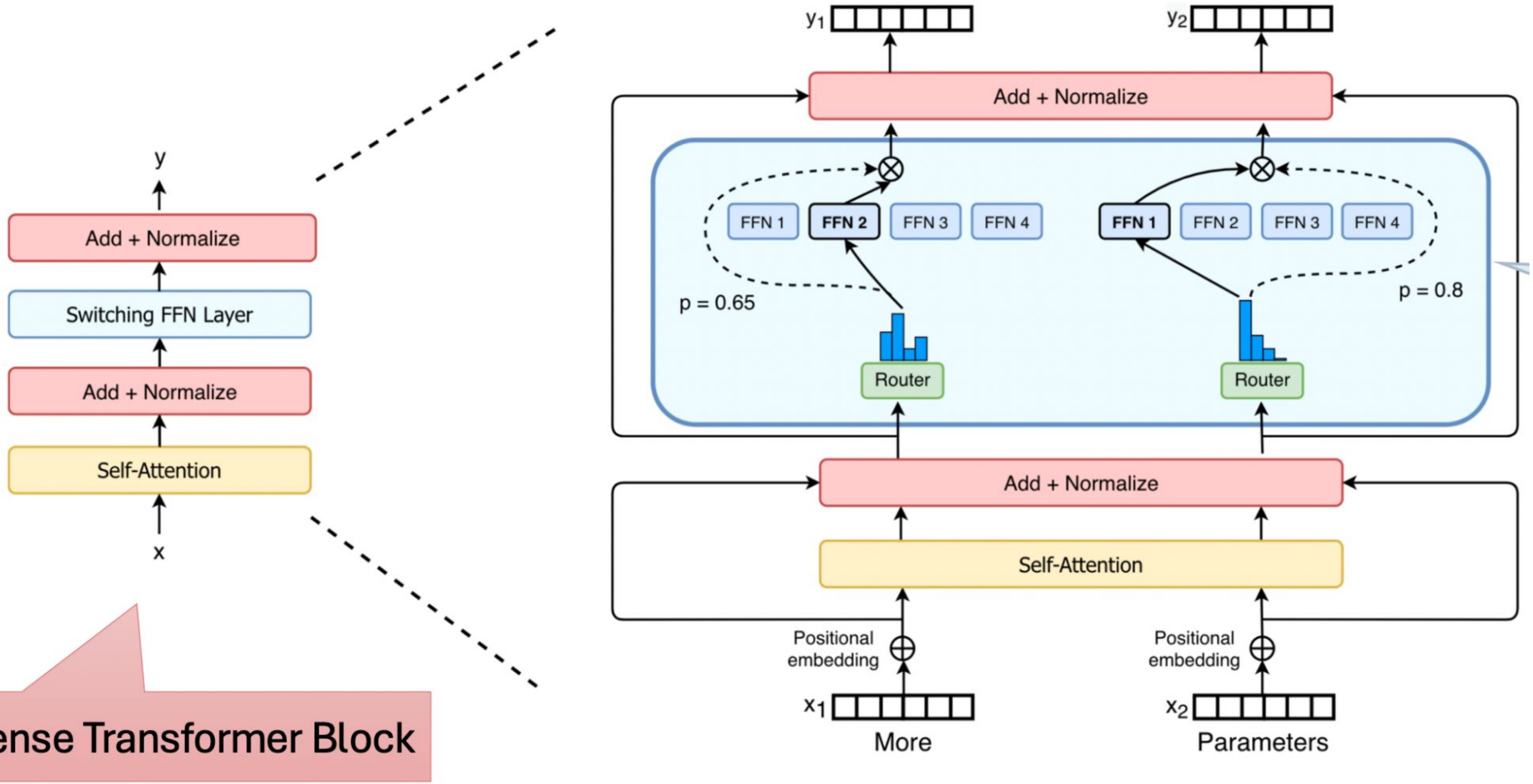
The Expert Mechanism

The magic of Mixtral lies in how it handles its feedforward block. Here's where the 'experts' come into play. Mixtral doesn't rely on a single set of parameters; instead, it picks from eight distinct groups of parameters. This selection is dynamic and context-dependent.

- **Token Routing**: For every token in the input, a router network chooses two groups of experts. This dual selection allows for a nuanced and context-rich processing of information.
- **Additive Output Combination**: The outputs from these chosen experts are then combined additively, ensuring a rich blend of specialized knowledge.


MoE layer
+
x
x
G(x)$_2$
G(x)$_{n-1}$
Expert 1
Expert 2
Expert 3
· · ·
Expert n-1
Expert n
Gating
Network
x

Mixture of Experts as a Layer

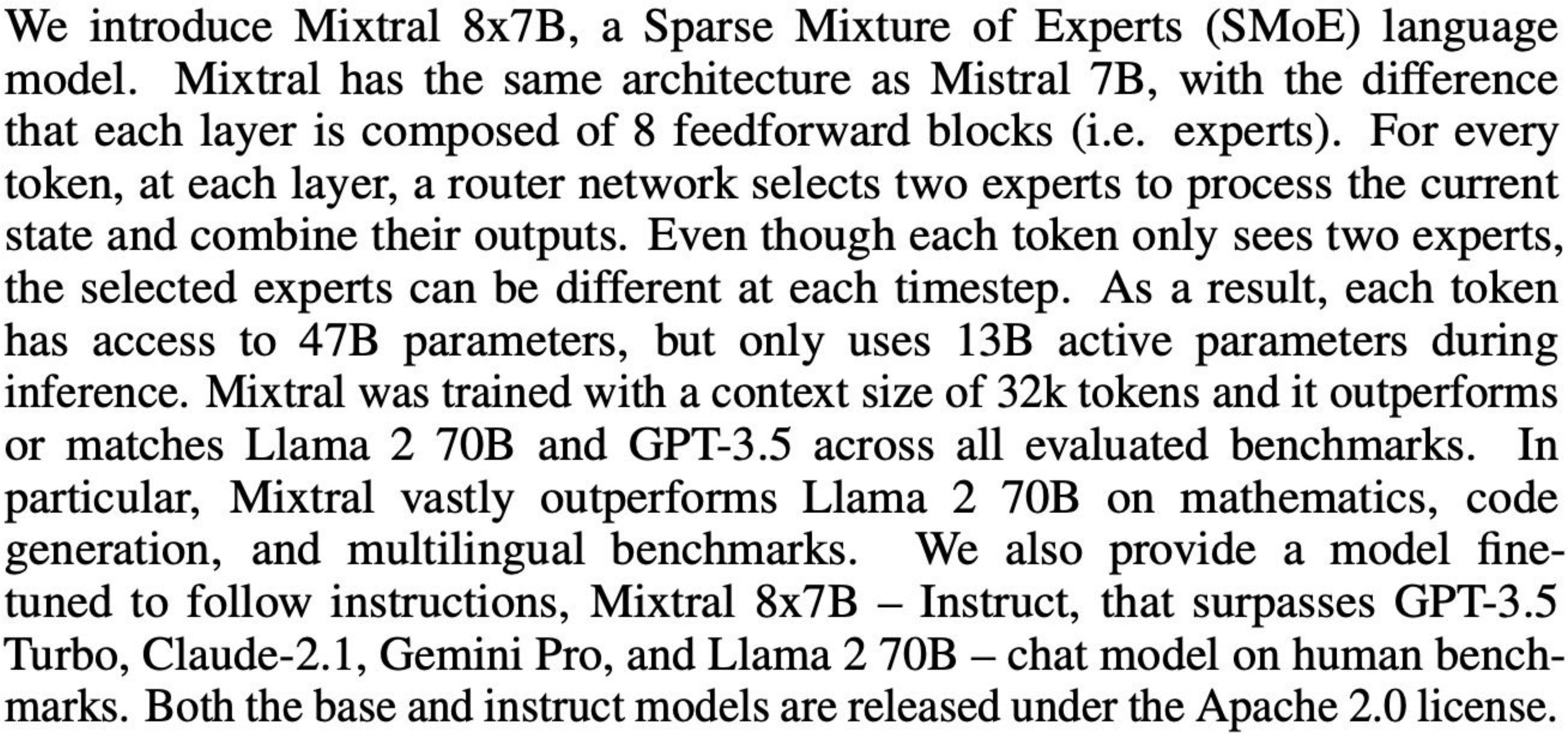


Dense Transformer Block

Mixtral of Experts

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch,
Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas,
Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour,
Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux,
Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao,
Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, William El Sayed

Mistral.AI



We introduce Mixtral 8x7B, a Sparse Mixture of Experts (SMoE) language model. Mixtral has the same architecture as Mistral 7B, with the difference that each layer is composed of 8 feedforward blocks (i.e. experts). For every token, at each layer, a router network selects two experts to process the current state and combine their outputs. Even though each token only sees two experts, the selected experts can be different at each timestep. As a result, each token has access to 47B parameters, but only uses 13B active parameters during inference. Mixtral was trained with a context size of 32k tokens and it outperforms or matches Llama 2 70B and GPT-3.5 across all evaluated benchmarks. In particular, Mixtral vastly outperforms Llama 2 70B on mathematics, code generation, and multilingual benchmarks. We also provide a model fine-tuned to follow instructions, Mixtral 8x7B – Instruct, that surpasses GPT-3.5 Turbo, Claude-2.1, Gemini Pro, and Llama 2 70B – chat model on human benchmarks. Both the base and instruct models are released under the Apache 2.0 license.

- Mixtral is a sparse mixture-of-experts network.

- It is a decoder-only model where the feedforward block picks from a set of 8 distinct groups of parameters.

- At every layer, for every token, a router network chooses two of these groups (the “experts”) to process the token and combine their output additively.

- This technique increases the number of parameters of a model while controlling cost and latency, as the model only uses a fraction of the total set of parameters per token.

Advantages of MoE

- Faster inference

- Lower Costs

- Quality of Answers

Disadvantages of MoE

- GPU VRAM Requirements

- Instabilities in pre-training

- Difficulties in fine-tuning

- Router Collapse