# Research Report

*Generated: 2024-11-15 17:14:54*

## Research Report

### Key Points

- **Introduction of Mixtral 8x7B**: The Mixtral 8x7B is a Sparse Mixture of Experts (SMoE) language model designed to enhance the efficiency and performance of language processing tasks. It represents a significant advancement in the architecture of language models, allowing for more effective utilization of parameters.
- **Architecture**: The architecture of Mixtral consists of 8 feedforward blocks, referred to as experts, per layer. This design enables the model to dynamically select which experts to engage for processing each token, thereby optimizing computational resources.
- **Parameter Access**: Each token processed by the model has access to a total of 47 billion parameters. However, during inference, the model only activates 13 billion parameters. This selective activation allows for efficient processing while maintaining high performance.
- **Training Context**: Mixtral was trained with a context size of 32,000 tokens, which is significantly larger than many existing models. This extensive context allows the model to better understand and generate language, leading to improved performance across various tasks.
- **Performance Benchmarking**: The Mixtral model has been shown to outperform or match the performance of other leading models, such as Llama 2 70B and GPT-3.5, across all evaluated benchmarks. Notably, it excels in areas such as mathematics, code generation, and multilingual tasks, demonstrating its versatility and capability.
- **Instruct Model**: In addition to the base model, the researchers have developed a fine-tuned version known as Mixtral 8x7B - Instruct. This variant is specifically designed to follow instructions and has been shown to surpass other models, including GPT-3.5 Turbo, Claude-2.1, Gemini Pro, and Llama 2 70B - chat model, on human benchmarks. This indicates its effectiveness in practical applications where instruction-following is critical.

### References and Sources

- Summary: 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.