**Assignment 04: Mistral vs Mixtral**
**Student: Lam Nguyen**
**Class: CAP 6411**

We will be discussing the Mistral and Mixtral Large Language Models. These are LLMs developed by Mixtral intended for Natural Language Processing Tasks like creating text, summarization and conversational AI.

## Mistral

### Architectural Details of Mistral-7B

**1.** Sliding Window Attention: Trained with an 8K context length and a fixed cache size, Mistral-7B offers a theoretical attention span of 128K Tokens.

**2.** Grouped Query Attention (GQA): Enables faster inference and reduces cache size, enhancing efficiency.

**3.** Byte fallback BPE Tokenizer. Ensures that characters are never mapped to out-of-vocabulary tokens

## Mixtral

Mixtral is an advancement over Mistral. It introduces a sparse mixture of experts (SMoE) model with open weights. Mixtral-8x7B outperforms Mistral and offers 6x faster inference.

### Architectural Details Mixtral-8x7B

**1.** Mixture of Experts (MoE) Model: 8 experts per MLP, Mixtral achieves its performance with open weights
**2.** Use flash attention: Speeds up inference by optimizing the attention mechanism
**3.** Quantization: Mixtral can be quantized to reduce memory footprint, enabling efficient deployment on consumer hardware

## Results:

In order to get the Mistral and Mixtral to run, I had to perform 4-bit quantization. The Mistral model took up about 4.3GB of VRAM and the Mixtral model took about 27GB of VRAM. The Evaluation Function was borrowed from the paper “Measuring Massive Multitask Language Understanding.” Unfortunately, for my case, I wasn’t able to get the Evaluation function working in time. And before I was able to solve the bugs, I ran out of Computing Time using Google Colab. However in theory, the Mixtral model would outperform the Mistral one. The code is included with this report.

## Source:

**Intro:** https://medium.com/@harshaldharpure/understanding-mistral-and-mixtral-advanced-language-models-in-natural-language-processing-f2d0d154e4b1

**Evaluation Pipeline:** https://github.com/hendrycks/test?tab=readme-ov-file