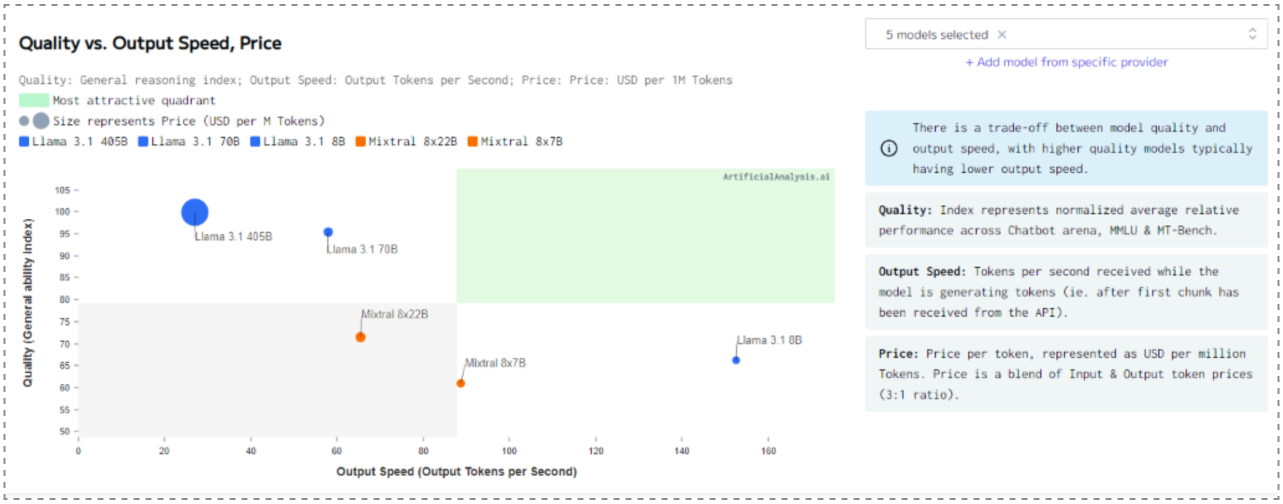
LLM, RAG – Llama Alternative

Hacker GPT Model

Mixtral 8X22B is a Large Language Model(LLM) developed by Mistral AI that is designed to handle NLP tasks such as text generation, summarization, and conversational AI.

Advantages of Mixtral:

- Multilingual capabilities; fluent in English, Spanish, French, German and Italian.
- Offers a sparse mixture of experts(SoME) architecture that reduces computational cost during pre-training.



Quality vs. Output Speed, Price

There is a trade-off between model quality and output speed, with higher quality models typically having lower output speed.

Quality: Index represents normalized average relative performance across Chatbot arena, MMLU & MT-Bench.

Output Speed: Tokens per second received while the model is generating tokens (ie. after first chunk has been received from the API).

Price: Price per token, represented as USD per million Tokens. Price is a blend of Input & Output token prices (3:1 ratio).

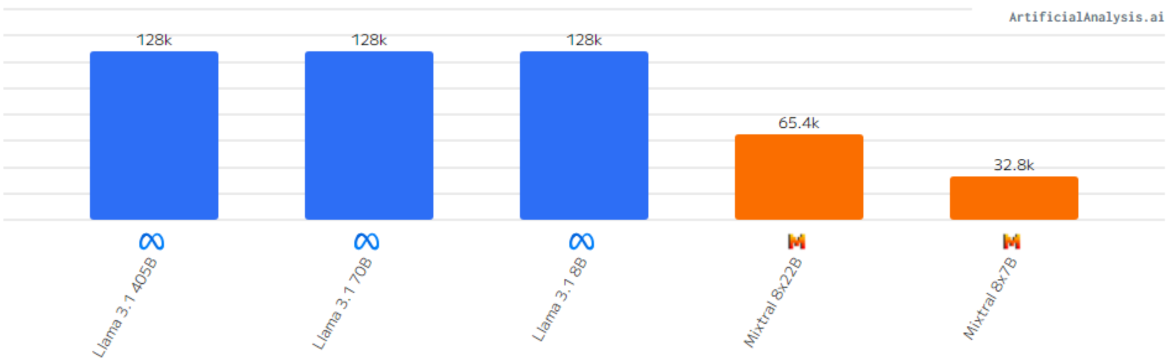
5 models selected ×
+ Add model from specific provider
Quality: General reasoning index; Output Speed: Output Tokens per Second; Price: Price: USD per 1M Tokens
Most attractive quadrant
Size represents Price (USD per M Tokens)
Llama 3.1 405B
Llama 3.1 70B
Llama 3.1 8B
Mixtral 8x22B
Mixtral 8x7B
ArtificialAnalysis.ai
Quality (General ability Index)
105
100
95
90
85
80
75
70
65
60
55
50
Llama 3.1 405B
Llama 3.1 70B
Mixtral 8x22B
Mixtral 8x7B
Llama 3.1 8B
0
20
40
60
80
100
120
140
160
Output Speed (Output Tokens per Second)

Disadvantages of Mixtral 8X7B compared to Llama 3.1 8B

- Output speed is slower than Llama 3.1 8B.
- Both Mixtral models costs more than Llama 3.1 8B per 1 million tokens.
- Token limit is smaller for Mixtral 8X22B model than any Llama 3 model.
- Requires more RAM vs the Llama 3.1 8B model. 16GB minimum vs 64GB minimum of RAM for optimal performance.
- When scaling up to try and improve accuracy, resource requirement of a 300 GB GPU running Mixtral 8X22B on cloud and consumer software make it very expensive and challenging.

Context window

Context window: Tokens limit; Higher is better


ArtificialAnalysis.ai
128k
128k
128k
65.4k
32.8k
Llama 3.1 405B
Llama 3.1 70B
Llama 3.1 8B
Mixtral 8x22B
Mixtral 8x7B