# Ollama Models Comparison

* The models have been tested on AWS GPU series EC2.
* The CPU models has been tested on AWS Ec2 of 32 GB RAM.
* The CPU\_GPU models have been tested on AWS Ec2 of 16GB RAM, 16GB GPU memory.
* The RAG response for RAG in CPU is on average 180 seconds, for features it is 15 seconds.
* The RAG response for RAG with GPU is on average 7 seconds, for features it is 3 seconds.

The models have been tested for

1. RAG, Entity recognition,
2. Sentiment analysis,
3. Reasoning ,
4. Analytical RAG, Summary.

For the above headings , the winner is mistral-nemo, runner up being llama 3.1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **RAG** | **Query complexity** | **Versatility** | **Definitive** | **Size** |
| Mistral | Good | Good | Average | Average | 4.5GB |
| Gemma2 9B | Better | Good | Good | Good | 5.4GB |
| Llama 3.1 8B | Better | Better | Better | Good | 4.7GB |
| Mistral -Nemo | Best | Best | Best | Good | 7.1GB |

**Conclusion**:

For offline model deployment such as Knowledge assistant, Chatbots, mistral-nemo stands the best choice.