**LLM:**

<https://huggingface.co/spaces/ArtificialAnalysis/LLM-Performance-Leaderboard>

AWS provided API for

**LLaMa 3.2 3B Instruct** (us.meta.llama3-2-3b-instruct-v1:0)

<https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct>

* Context window: 128K
* Cost/1M tokens: $0.15
* Supported use cases: Advanced text generation, summarization, sentiment analysis, emotional intelligence, contextual understanding, and common sense reasoning
* Knowledge Cutoff: **December 2023**

**Mistral 7B Instruct** (mistral.mistral-7b-instruct)

<https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2>

<https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3>

* Context window: 33K
* Cost/1M tokens: $0.16
* Supported use cases: Text summarization, structuration, question answering,  
  and code completion
* Knowledge Cutoff: **December 2023**

**Vector DB:**

<https://zilliz.com/comparison/chroma-vs-faiss>

**Faiss DB** (created by Meta)

* Enables swift search of similarities and the clustering of dense vectors
* It houses algorithms capable of searching within vector sets of varying sizes, even those that might exceed RAM capacity
* Some of its key algorithms are also available for GPU execution

**Chroma DB**

* Feature-rich: queries, filtering, density estimates, and many other features
* [LangChain](https://www.datacamp.com/tutorial/how-to-build-llm-applications-with-langchain) (Python and JavScript), [LlamaIndex](https://www.datacamp.com/tutorial/llama-index-adding-personal-data-to-llms" \t "_blank), support available
* The same API that runs in Python notebook scales to the production cluster

**SQL DB:**

<https://www.geeksforgeeks.org/difference-between-sqlite-and-postgresql/>

|  |  |
| --- | --- |
| **FAISS** | **ChromaDB** |
|  | Embeddings Stored along with Meta Data. Would work well for DBs |
| Fast | Fast |
| Works well with Various Dtypes | Works well with High Dimension Data |
| Optimized Data storage |
| Resource Intensive | Lightweight |
| Image Systems | High Dimension systems |
| Scaling can be challenging | Easier to Scale |
| Provides GPU Support | Can host on Local GPUs |