Phase - 1 : Research

Libraries & Selection of Local LLM for M.I.K.E

Models :

DeepSeek-R1(8B) :

- Parameters: 8 billion
- Architecture: Transformer-based
- Purpose: NLP tasks (generation, translation, summarization)
- Performance: High accuracy, large-scale processing
- Deployment: Versatile (chatbots, QA systems.

- Mistral - 7b
  - Parameters: 7 billion
  - Architecture: Transformer-based
  - Purpose: NLP tasks (text generation, summarization, question answering)
  - Performance: Efficient and scalable for large tasks
  - Deployment: Suitable for real-time applications and APIs

Model Comparison:

We tested both models with the below questions :

- Basic: Capital of France?
- Intermediate: Differences between supervised and unsupervised learning?
- High-Level: Designing a bias-aware ML pipeline for customer purchases.

Feature	DeepSeek-R1	Mistral Large
Context Window	64,000 tokens	32,000 tokens
MMLU Score	90.8%	81.2%
Advantages	- Larger context window for better comprehension of long documents - Higher benchmark performance (MMLU: 90.8%)	- More established with community support - More resource-efficient, faster for shorter tasks
Disadvantages	- Higher resource consumption - Newer, with potentially fewer integrations	- Smaller context window limits long document handling - Lower benchmark performance than DeepSeek-R1

Model Overview

Feature	Mistral Large	DeepSeek-V3
Input Context Window The number of tokens supported by the input context window.	**32K** tokens	**128K** tokens
Maximum Output Tokens The number of tokens that can be generated by the model in a single request.	**4,096** tokens	**8K** tokens
Open Source Whether the model's code is available for public use.	**Yes**	**Yes**
Release Date When the model was first released.	**February 26, 2024** 11 months ago	**December 27, 2024** 1 month ago

Model Performance

Benchmark Comparison

Compare performance metrics between Mistral Large and DeepSeek-V3. See how each model performs on key benchmarks measuring reasoning, knowledge and capabilities.

Benchmark	Mistral Large	DeepSeek-V3
MMLU Massive Multitask Language Understanding - Tests knowledge across 57 subjects including mathematics, history, law, and more	**81.2%** 5-shot Source	**88.5%** EM Source
MMLU-Pro A more robust MMLU benchmark with harder, reasoning-focused questions, a larger choice set, and reduced prompt sensitivity	Not available	**75.9%** EM Source
MMMU Massive Multitask Multimodal Understanding - Tests understanding across text, images, audio, and video	Not available	Not available
HellaSwag A challenging sentence completion benchmark	**89.2%** 10-shot Source	**88.9%** 10-shot Source
HumanEval Evaluates code generation and problem-solving capabilities	Not available	**82.6%** pass@1 Source
MATH Tests mathematical problem-solving abilities across various difficulty levels	Not available	**61.6%** 4-shot Source
GPQA Tests PhD-level knowledge in chemistry, biology, and physics through multiple choice questions that require deep domain expertise	Not available	**59.1%** pass@1 Source
IFEval Tests model's ability to accurately follow explicit formatting instructions, generate appropriate outputs, and maintain consistent instruction adherence across different tasks	Not available	**86.1%** Prompt Strict Source

Results :

- Findings DeepSeek-R1 (8B) provided excessive details, while Mistral 7B delivered concise, relevant responses.

Decision : We are proceeding with Mistral 7B for its precision & efficiency.

< —---------- >

Tools & Libraries

- VSCode (IDE)
- Python
    - PyPDF2 (Extracts Text from PDF)
    - HuggingFace
    - Ollama (Run LLM's Locally)
    - Mistral (LLM)
    - Sentence Transformers (Convert text into embeddings)
        - mpnet_v2
        - distilroberta
    - Faiss-CPU (Efficient Text-Search)
    - Streamlit (GUI)