FOSDEM’24

Using Generative AI and Content Service Platforms together

Angel Borroy
Developer Evangelist

January 25, 2024
Agenda

- GenAI Stack
- Integration
  - Features
  - Existing Content
  - New Content
- Demo
- What Else?
GenAI Stack
Docker GenAI Stack

Components

ollama
- Local management of *open source* LLMs
- Catalog of preconfigured LLMs, such as Llama2 or Mistral

neo4j
- Graph and native vector search capabilities
- Ground LLMs for more precise GenAI predictions and outcomes

LangChain
- Communication between the LLM, your application, and the database
- Python framework for developing applications powered by LLMs
Docker GenAI Deployment

Deployment

https://github.com/docker/genai-stack

- pull-model
- ollama
- neo4j
- streamlit
- LangChain
- REST API
- GPU
Docker GenAI

Meta
MISTRAL AI
TII
deepseek coder
Microsoft
deepseek coder

Llama2 Community License Agreement
Llama2 Community License Agreement
Apache License 2.0
Deepseek License Agreement
Llama2 Community License Agreement
Falcon TII License 1.0

llama2
codellama
mixtral
deeplake-coder
vicuna
falcon

13+B

40+ GB
Size
4
64+ GB
RAM
1024
1024

phi
dolphin-phi
orca-mini
deepseek-coder

3B

4 GB
RAM
8 GB
Size
8 GB
16 GB

MIT License
cc-by-nc-sa-4.0
Deepseek License Agreement

Llama2 Community License Agreement
Llama2 Community License Agreement
Llama2 Community License Agreement
Apache License 2.0
Apache License 2.0
Apache License 2.0
cc-by-nc-sa-4.0
Deepseek License Agreement
Microsoft Research License
Microsoft Research License
Microsoft Research License
Falcon TII License 1.0

Llama2 Community License Agreement
Llama2 Community License Agreement
Llama2 Community License Agreement
Apache License 2.0
Apache License 2.0
Apache License 2.0
cc-by-nc-sa-4.0
Deepseek License Agreement
Microsoft Research License
Microsoft Research License
Microsoft Research License
Falcon TII License 1.0
Docker GenAI

Ethical AI Rating

Mistral 7B

Is the software (both for inferencing and training) open source?
https://github.com/mistralai/mistral-src

Is the trained model freely available for self-hosting?
https://ollama.ai/library/mistral

Is the training data available and free to use?
No
Integration
Integration with Content Service Platform

Deployment

**ALFRESCO**
- model-repo
- model-share

**GENAI STACK**
- llm
- neo4j
- langchain

**Docker Network**
- tcp://activemq:61616
- http://alfresco:8080
- http://genai:8506/summary
- http://genai:8506/classify
- http://genai:8506/prompt

https://github.com/aborroy/alfresco-genai
Alfresco Integration

GenAI App

Python application that provides REST API endpoints

- **Document Loader**: Split document into smaller chunks
- **Store in VectorDB**: Embed and save chunks
- **Create QA Chain**: With the ability to lookup Vector DB

- **LLM Embeddings**
- **LLM**

**SUMMARY**:

- **LLM**: mistral
- **SUMMARY_LANGUAGE**: English
- **SUMMARY_SIZE**: 120
- **TAGS_NUMBER**: 3
Demo

MacBook Pro 2021
Apple M1 Pro
32 GiB RAM

GPU
curl --location 'http://localhost:8506/summary' --form 'file=@"./file.pdf"'
{
    "summary": "The text discusses...",
    "tags": "Golang, Merkle, Difficulty",
    "model": "mistral"
}

curl --location \\
--form 'file=./file.pdf'
{
    "term": "English",
    "model": "mistral"
}
curl --location 'http://localhost:8506/prompt?prompt="What is this text about?"' \
--form 'file=./file.pdf''
{
  "answer": "Yes, it is difficult to find childcare in Tokyo.",
  "model": "mistral"
}
What Else?
Find your way

Docker AI/ML Hack-a-thon

- Readme AI
- Techdocs
- Docker Image Analyzer
- Docker Log Sentiment Analyzer
- GitChats AI

ollama alternatives

- https://gpt4all.io/index.html
- https://localai.io
- https://www.secondstate.io/run-llm
- https://huggingface.co/docs/hub/spaces-sdks-docker-first-demo
Thanks!

Hyland