

# Leveraging MLflow for Efficient Deployment and Evaluation of Large Language Models

*Lisana Berberi, Khadijeh Alibabaei, Borja Esteban Sanchis, Valentin Kozlov*

*Karlsruhe Institute of Technology*



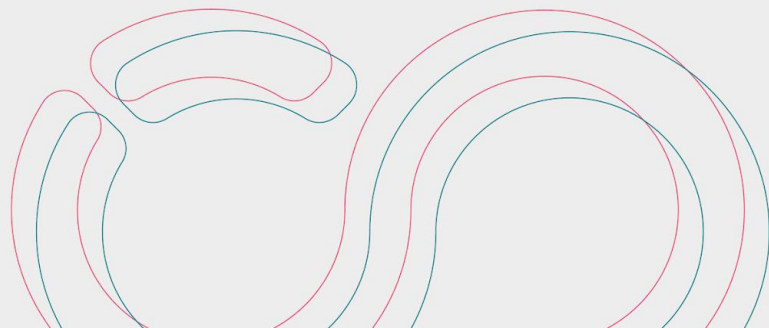
Funded by  
the European Union

2024 | 10 | 03 by Lisana Berberi



# Outline

- Introduction on LLM
- MLflow Evaluation and Deployment
- Open LLM Comparison
- Conclusions and Outlook



# Large Language Models (LLMs)

- Advanced machine learning model designed to
  - **understand**,
  - **generate**,
  - **manipulate** natural language with high accuracy
- Are built using deep learning techniques, specifically using neural network architectures like **transformers** [1]
- Utilized in applications:
  - **question-answering**,
  - content generation,
  - summarization, and
  - decision-making



# Key Characteristics of LLMs

- **Scale and Size:** LLMs are characterized by their large number of parameters, often ranging from one billion to hundred of billions.
  - This allows them to capture complex language patterns and nuances.
- **Training Data:** They are trained on vast amounts of text (audio, video and image) data from diverse sources such as books, articles, and websites.
  - This extensive training enables them to have a broad understanding of language.

## Applications:

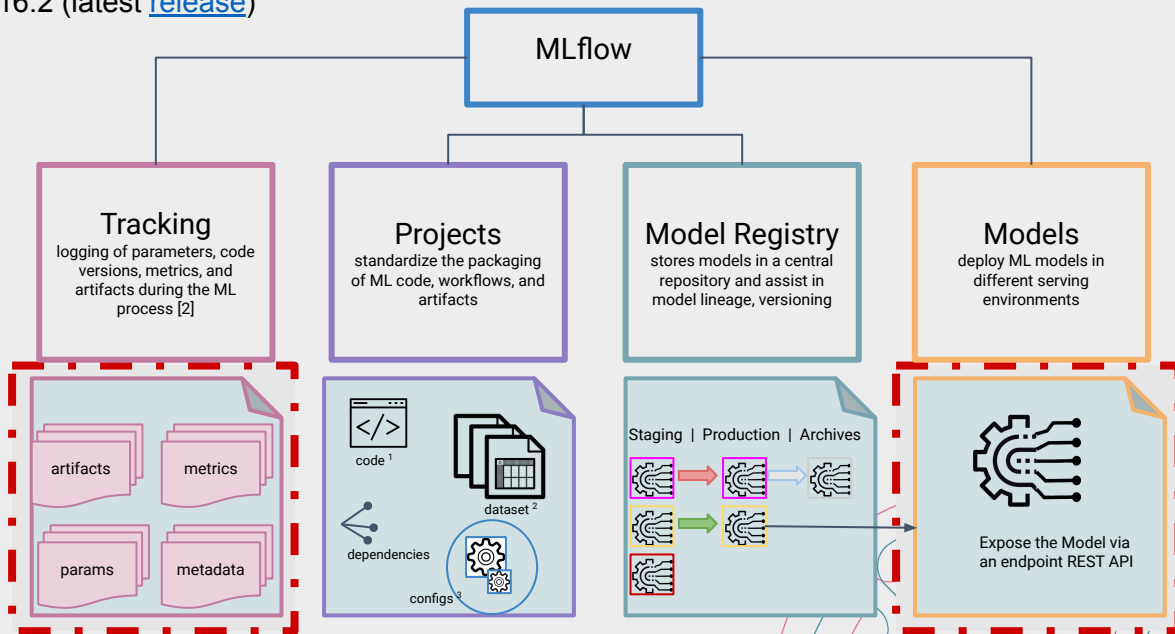
- **Chatbots** and Virtual Assistants: Powering intelligent conversational agents.
- **Content Creation:** Assisting in writing articles, blogs, and other content.
- **Code Generation:** Generating and suggesting code snippets for programming tasks.
- **Research and Education:** Assisting researchers and students with information retrieval and summarization.

## Examples:

- LLaMA (Large Language Model Meta AI) by Meta.
- GPT (**Generative** Pre-trained Transformer) models by OpenAI (e.g., GPT-3, GPT-4).
- BERT (Bidirectional Encoder Representations from Transformers) by Google.



- An open source platform for the machine learning lifecycle
- mlflow 2.16.2 (latest [release](#))



MLFlow Components



Funded by  
the European Union

2024 | 10 | 03 by Lisana Berberi

<sup>1</sup> Code icon by [zajour mohcne](#) licensed under the [CC BY 3.0 license](#)  
<sup>2</sup> Data set icon by [H.Alberto Góngora](#) under the license [CC BY 3.0 license](#)  
<sup>3</sup> Config icon by [Madalin Jefferson](#) under the license [CC BY 3.0 license](#)

# Why MLflow for LLMs?

- **Resource-Efficient Evaluation with MLflow**

Using MLflow to assess model performance on specific tasks without extensively deploying the model in production:

- **Lightweight Evaluation:** MLflow's `evaluate` function allows for quick performance assessments using various metrics (e.g., accuracy, BLEU score) on a subset of data.
  - This pre-deployment evaluation provides insights into the model's performance without the need for extensive computational resources.
- **Feedback Mechanism:** By logging predictions and comparing them against ground truth labels, MLflow helps identify areas where the model might need improvement (e.g., fine-tuning or additional training data).
  - This feedback loop enables iterative model refinement in a cost-effective manner.
- **Avoids Costly Re-runs:** Instead of deploying models in full-scale environments, which is resource-intensive, MLflow can evaluate models on sampled datasets.
  - This provides preliminary feedback on model performance, reducing the need for large-scale tests and fine-tuning iterations.
- **Deploy LLM locally**

```
mlflow models serve -m {mlflow-artifacts-uri} --port {port} --host {host}
```



```
mlflow.evaluate(  
    model_uri,  
    alpaca_data.head(10), # Adjust the number of rows as needed  
    targets="ground_truth",  
    extra_metrics=[  
        mlflow.metrics.ari_grade_level(),  
        mlflow.metrics.latency(),  
        mlflow.metrics.flesch_kincaid_grade_level()  
    ],  
)
```

**ari\_grade\_level()**= This metric outputs a number that approximates the grade level needed to comprehend the text, which will likely range from around 0 to 15

**latency()**= Latency is determined by the time it takes to generate a prediction for a given input.

**flesch\_kincaid\_grade\_level()**= This metric outputs a number that approximates the grade level needed to comprehend the text, which will likely range from around 0 to 15 [formula diff: includes syllable]

# How to evaluate LLMs with MLflow?

- **MLflow AI Gateway (Experimental)**

The MLflow AI Gateway service is a powerful tool designed to streamline the usage and management of various large language model (LLM) providers. It offers a high-level interface that simplifies the interaction with these services by providing a **unified endpoint** to handle specific LLM related requests.

```
# Start the Gateway Service
```

```
mlflow gateway start --config-path config.yaml --port {port} --host {host} --workers {worker count}
```

- **MLflow's LLM evaluation functionality consists of 3 main components:**
  - **A model to evaluate:** it can be an MLflow `pyfunc` model, a URI pointing to one registered MLflow model, or any python callable that represents your model, e.g, a HuggingFace text generation pipeline.
  - **Metrics:** the metrics to compute, LLM evaluate will use LLM metrics.
    - `ari_grade_level()`, etc.
  - **Evaluation data:** the data your model is evaluated at, it can be a pandas Dataframe, a python list, a numpy array or an [mlflow.data.dataset.Dataset\(\)](#) instance.
- **Our setup:** 2 VMS, Tesla T4 GPU:16GB, disk: 256GB, CPU RAM:43GB





## Open LLM

Features	Parameter	Inference	Architecture	Time (GPU hours) /Power Consumption (W) C/ Emitted(tCO <sub>2</sub> eq)	Trained date	License
Mistral (Mistral-7B-Instruct-v0.1)[ 2,3]	7B	Yes	-Grouped-Query Attention -Sliding-Window Attention -Byte-fallback BPE tokenizer	Not explicitly mentioned	September 2023	Apache license 2.0
Llama (Llama-2-7B-Chat-hf) [4,5]	6.7B	Yes	-The tuned versions use supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align to human preferences for helpfulness and safety.	184320/ 400/ 31.22	between January 2023 and July 2023.	Llama 2 Community License Agreement





# Example of a prompt execution using AI Gateway

### New run

Create a new run using a large-language model by giving it a prompt template and model parameters

**Served LLM model**

Select LLM model endpoint

- chat-llama  
llama-2-7b
- chat-mistral  
mistral-7b

**Prompt Template** [View Examples](#)

Give instructions to the model. Use {{ }} or the "Add new variable" button to add variables to your prompt.

{{ var }}

**var**

What is MLflow?

+ Add new variable

▶ Evaluate

**Output**

This is the output generated by the LLM using the prompt template and input values defined above.

Cancel Create run

0 matching runs



# Example of two prompt executions using AI Gateway

### New run

Create a new run using a large-language model by giving it a prompt template and model parameters

**Model parameters**

Temperature  ⓘ

0.01

Max tokens  ⓘ

Stop Sequences  ⓘ

**New run name**

**Prompt Template**

**var**

+ Add new variable

▶ Evaluate

**Output**

This is the output generated by the LLM using the prompt template and input values defined above.

MLflow is an open-source platform for the end-to-end management of machine learning experiments. It provides a framework for tracking experiments, packaging code and data, sharing models, and deploying models to production. MLflow is designed to help teams manage the complexity of machine learning projects and improve collaboration and reproducibility.

Cancel
Create run

0 matching runs

### New run

Create a new run using a large-language model by giving it a prompt template and model parameters

**Served LLM model**

**Model parameters**

Temperature  ⓘ

0.01

Max tokens  ⓘ

Stop Sequences  ⓘ

**New run name**

**Prompt Template**

Give instructions to the model. Use {{ }} or the "Add new variable" button to add variables to your prompt.

View Examples

**var**

+ Add new variable

▶ Evaluate

**Output**

This is the output generated by the LLM using the prompt template and input values defined above.

MLflow is an open-source platform for managing the end-to-end machine learning (ML) lifecycle. It provides a suite of tools for versioning, reproducibility, and deployment of ML models. MLflow helps data scientists and engineers to manage the entire ML workflow, from data preparation and model training to deployment and monitoring of models in production environments.

Cancel
Create run

4 matching runs



## New run

Create a new run using a large-language model by giving it a prompt template and model parameters

## Model parameters

Temperature ⓘ

0.01

Max tokens ⓘ

256

Stop Sequences ⓘ

## New run name

overjoyed-bat-687

var




## Output

This is the output generated by the LLM using the prompt template and input values defined above.

Spark is an open-source data processing engine that can handle large-scale data processing and analytics. It is designed to be fast, scalable, and fault-tolerant, and can be used for a wide range of data processing tasks, including batch processing, streaming processing, and machine learning. Spark is written in Scala and can also be used with other programming languages, such as Python and Java.



0 matching runs

## New run

Create a new run using a large-language model by giving it a prompt template and model parameters

## Model parameters

Temperature ⓘ

0.01

Max tokens ⓘ

256

Stop Sequences ⓘ

## New run name

traveling-horse-301

var




## Output

This is the output generated by the LLM using the prompt template and input values defined above.

Spark is an open-source data processing engine that is designed to handle large-scale data processing tasks. It was developed at the University of California, Berkeley and is now maintained by Apache. Spark is a fast and flexible engine that can handle a wide variety of data processing tasks, including batch processing, stream processing, and machine learning.

Spark is written in Scala, but it also has APIs in Java, Python, and R. It can run on a variety of platforms, including Apache Hadoop, Apache Mesos, and standalone machines. Spark is highly scalable and can handle large amounts of data, making it a popular choice for big data processing tasks.



5 matching runs



Funded by  
the European Union

2024 | 10 | 03 by Lisana Berberi

# Inputs given from Alpaca and Mbpp datasets

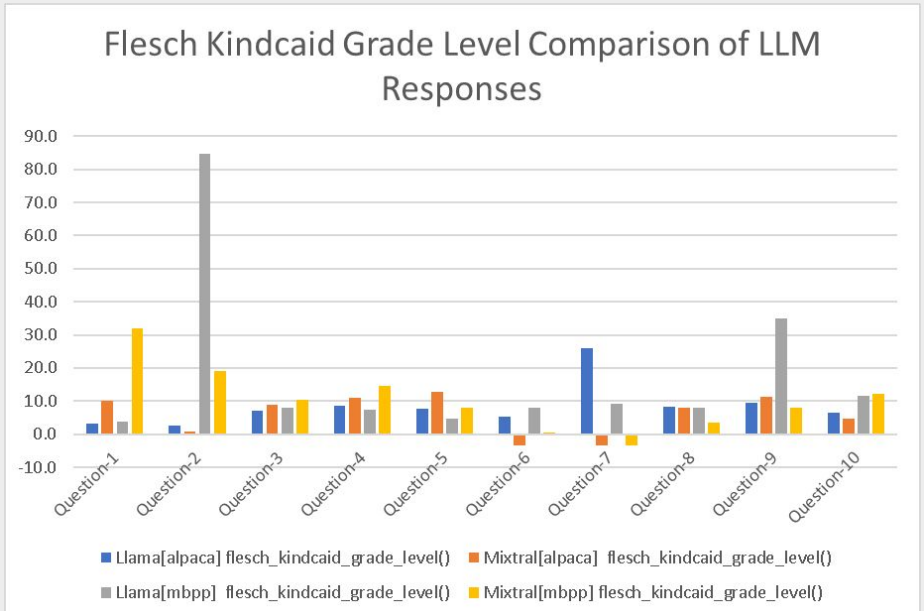
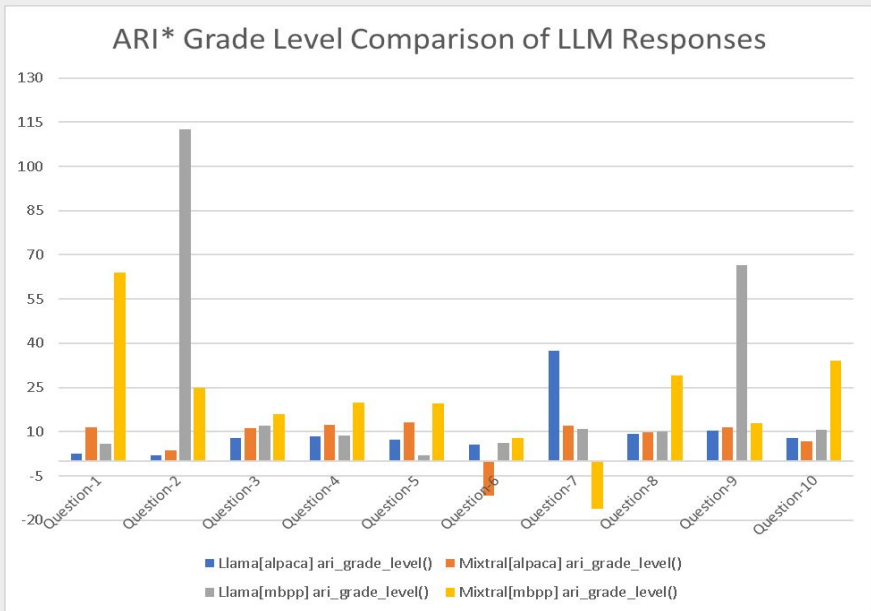
## Alpaca dataset

Inputs	Ground_truth	Outputs
(Question-1): Give three tips for staying healthy.	<ol style="list-style-type: none"><li>1. Eat a balanced diet and make sure to include plenty of fruits and vegetables.</li><li>2. Exercise regularly to keep your body active and strong.</li><li>3. Get enough sleep and maintain a consistent sleep schedule.</li></ol>	<ol style="list-style-type: none"><li>1. Eat a balanced diet that includes plenty of fruits, vegetables, lean proteins, and whole grains.</li><li>2. Exercise regularly, aiming for at least 150 minutes of moderate-intensity aerobic activity or 75 minutes of vigorous-intensity activity each week.</li><li>3. Get enough sleep, aiming for 7-9 hours per night. Prioritize sleep and try to maintain a consistent sleep schedule.</li></ol>

## MBPP dataset

Inputs	Ground_truth	Outputs
(Question-10): Write a python function to find the volume of a triangular prism.	<pre>def find_Volume(l,b,h) :     return ((l * b * h) / 2)</pre>	<pre>def volume_triangular_prism(length, width, height, base_length, base_width):     volume = length * width * height     lateral_area = (base_length * height + base_width * length) / 2     top_area = base_length * base_width     return volume + lateral_area - top_area</pre>

# Comparison charts Mixtral vs Llama



Dataset 1: <https://huggingface.co/datasets/tatsu-lab/alpaca>  
 Dataset 2: <https://huggingface.co/datasets/Muennighoff/mbpp> (python code)



Funded by  
the European Union

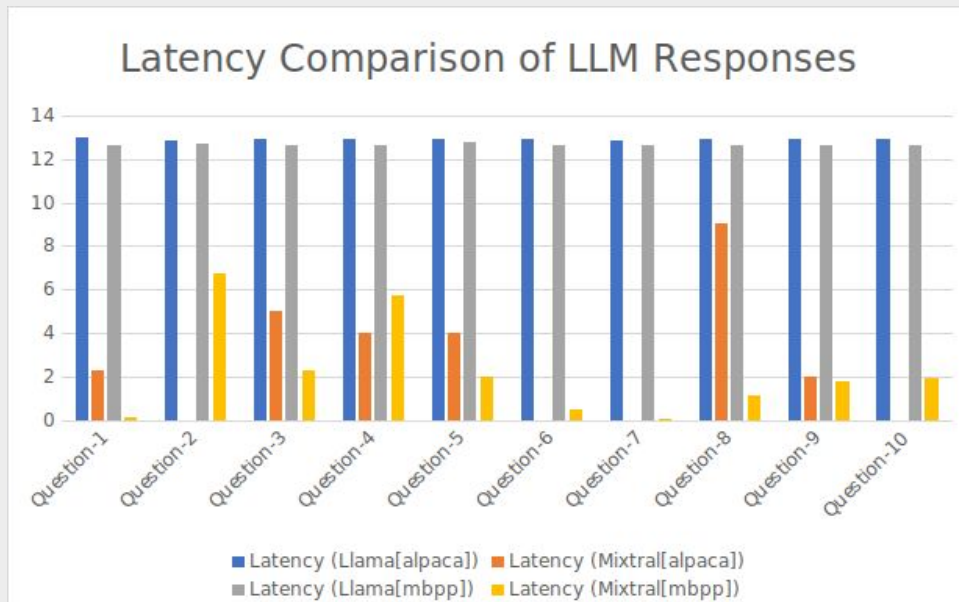
\*Automated Readability Index

2024 | 10 | 03 by Lisana Berberi

**-lower-is-better**

Lower values mean **easier** readability, so they are not designed to measure the depth of **reasoning** or the **knowledge** represented in the text. These metrics assess readability in terms of sentence complexity, word length, and syllables.

# Comparison charts Mixtral vs Llama



Dataset 1: <https://huggingface.co/datasets/tatsu-lab/alpaca>  
 Dataset 2: <https://huggingface.co/datasets/Muennighoff/mbpp> (python code)






Funded by  
the European Union

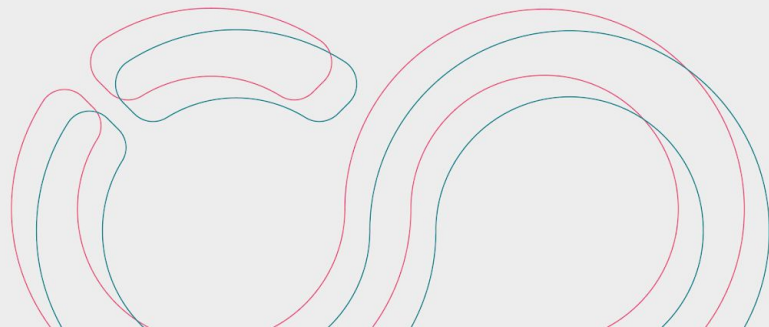
**\*Automated Readability Index**

2024 | 10 | 03 by Lisana Berberi



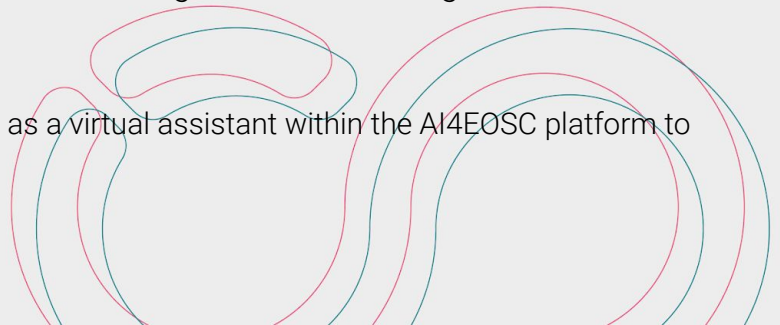
 Open LLM Leaderboard [7]

Model	Average 	IFEval	BBH	MATH Lv1 5	GPQA	MUSR	MMLU-PRO	Architecture	#Params (B)
<a href="#">mistralai/Mistral-7B-Instruct-v0.1</a> 	12.67	44.87	7.65	1.66	0	6.13	15.72	MistralForCausalLM	7
<a href="#">meta-llama/Llama-2-7b-chat-hf</a> 	9.4	39.86	4.46	0.68	0.45	3.28	7.64	LlamaForCausalLM	6



# Conclusions and Future Work

- **Advanced Model Development:**
  - LLMs are built using sophisticated deep learning techniques, specifically neural network architectures like transformers.
- **Evaluation with MLflow:**
  - Leverage MLflow functionalities to evaluate an open pre-trained LLM model on a specific dataset without deployment.
- **LLM as a Judge:**
  - Utilize the "LLM as a judge" (e.g. ChatGPT 4) methodology to assess model performance effectively.
- **Fine-Tuning Exploration:**
  - Investigate the possibility of fine-tuning one of the LLMs using the Retrieval-Augmented Generation (RAG) approach.
- **Virtual Assistant Deployment:**
  - Check if would be possible to deploy the fine-tuned model as a virtual assistant within the AI4EOOSC platform to provide guidance and support to users.





# References

1. Vaswani, A. "Attention is all you need." Advances in Neural Information Processing Systems (2017), <https://doi.org/10.48550/arXiv.1706.03762>
2. Huggingface Mistral-7b: <https://huggingface.co/mistralai/Mistral-7B-v0.1>
3. Mistral-7b paper: <https://arxiv.org/abs/2310.06825>
4. Huggingface LLama-2-7b-hf: <https://huggingface.co/meta-llama/Llama-2-7b-hf>
5. Llama paper: "Llama-2: Open Foundation and Fine-tuned Chat Models"
6. Source code: <https://codebase.helmholtz.cloud/m-team/ai/llm-evaluate/>
7. Fourier, C., Habib, N., Lozovskaya, A., Szafer, K., & Wolf, T. (2024). *Open LLM Leaderboard v2*. Hugging Face. [https://huggingface.co/spaces/open-llm-leaderboard/open\\_llm\\_leaderboard](https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard)



AI4

 eosc



Co-funded by  
the European Union



AI4EOSC



ai4eosc-po@listas.csic.es



ai4eosc.eu

# Reach us!

Thank you for your attention

Project Coordinator: Álvaro López García - [aloga@ifca.unican.es](mailto:aloga@ifca.unican.es)



Funded by  
the European Union