Max Eduardo Lazarini Wienandts

Project Topic: Mistral vs Llama2

Problem statement: Is Mistral better than Llama?

Abstract:

This study delves into the performance evaluation of Mistral and Llama2 LLMs, two prominent language models hailed for their advancements in natural language understanding. Despite Mistral's initial claims of superiority over Llama, this research aims to ascertain whether it genuinely outperforms its predecessor.

This study involves comparing both models using 11 distinct categories of questions or tasks. Initially, simple instructions were provided to assess the models' ability to complete the tasks. In the event of model failure or hallucination, prompt engineering techniques were applied to address the issue. If the model continued to face challenges in completing the task despite prompt engineering interventions, additional experiments were conducted exploring alternative inference parameters.

Results reveal a notable discrepancy between the models' capabilities. Llama exhibited competence across most categories without requiring prompt engineering. In contrast, Mistral consistently demanded prompt engineering and specific inference parameters across all tasks, indicating a higher degree of finetuning for optimal performance.

This comparative analysis sheds light on the nuanced performance dynamics between Mistral and Llama LLMs, offering valuable insights into their respective strengths, limitations, and applicability across diverse use cases.

The categories analyzed were:

1. Knowledge-based questions.
2. Reasoning questions.
3. Language understanding.
4. Inference questions.
5. Contextual understanding.
6. Creative tasks.
7. Ethical dilemmas.
8. Commonsense reasoning.
9. Translation tasks.
10. Summarization tasks.
11. Evaluation tasks.

Mistral was only able to correctly solve without any further instruction the categories 5, and 6.

Llama2 was able to correctly solve without any further instruction the categories 1, 2, 3, 4, 5, 6, 8, 10, and 11.

Youtube URL, short video: <https://youtu.be/c-0x3z-bO2M>

Youtube URL, full video: <https://youtu.be/QU1ZPKXKtM4>

1 – Introduction 3

2 - Brief introduction to Llama and Mistral 3

3 - Hardware and Software 5

4 – Python Code 5

5 - Methodology 7

6 - Results 9

7 - Conclusion 11

# 1 - Introduction

After the release of chatGPT-3 for the public by Open AI in November 2022, large language models (LLMs) gained visibility. Open AI’s model is easy to use and give straight answer for most of the questions. However, there is still several features that needs improvement. Therefore, several big tech companies started to invest in their owns models. Two of these companies are Meta and Mistral AI.

Meta was originally founded as TheFacebook in 2004 and is one of the Big Five American information technology company. Its first LLM released to the public was Llama (Large Language Model Meta AI) in February 2023. After, in July 2023, Meta released a new version Llama 2.

Mistral AI is a French company founded in April 2023, specialized in artificial intelligence products. Its first LLM released to the public was Mistral 7B in September 2023. The company announced that its model outperformed all Llama’s model in most of the metrics used to calculate a LLM performance (<https://mistral.ai/news/announcing-mistral-7b/>).

This study will demonstrate how to use both technologies, Llama 2 and Mistral, and compare its performance in 11 different categories of questions or instructions by judging how easy or hard it is to obtain the desired output. This study is not a confrontation or verification of the performance metrics released by Mistral AI. What it will be compared is whether any user, even the ones not familiarized with the technology, can fully use these tools without the need of prompt engineering or any other more advanced technique.

First, I will give a brief introduction of Llama 2 and Mistral models. After, I will specify the hardware and software used to conduct this experiment. Following, I give further instructions about how to reproduce the Python code used. Next, I will specify the questions/instruction categories tested and the techniques used to achieve a proper result when necessary. Finally, I will present the results and finish this study with a conclusion.

Together with this report you can find:

- “Mistral\_vs\_Llama\_WienandtsMax\_onePage.docx”: one page file containing the project summary.

- “Mistral\_vs\_Llama\_WienandtsMax.pptx”: project presentation.

- ‘’ e104\_project.yml”: YLM file with the Python’s libraries needed to run this study.

- “Llama.ipynb”: Jupyter notebook with the questions/instructions to test Llama 2 model.

- “Mistral.ipynb”: Jupyter notebook with the questions/instructions to test Mistral model.

# 2 - Brief introduction to Llama and Mistral

Llama was release in February 2023 by Meta. This model was trained with 1.4 trillion tokens and is available in 4 different sizes: 7B, 13B, 33B, and 65B of parameters. For comparison, GPT-3 model has 175B parameters.

After, in Jully 2023, Meta in a partnership with Microsoft launched Llama 2. This new version was trained with 40% more data (2 trillion tokens), but is available in only 3 sizes: 7B, 13B, and 70B parameters.

Recently, on April 18, 2024, Meta launched Llama 3. This model was trained with 15 trillion tokens and is available in only 2 sizes: 8B and 70B parameters. Unfortunately, as of April 26, 2024, Llama 3 is not available to use in Brazil (if using the web site <https://llama.meta.com/llama3/>), and its use by Huggin Face needs an authorization.

Mistral 7B was released in September 2024. An official number of trained tokens could not be found in Mistral AI website. As the implicit in its name, this model has 7 billion parameters. When released, Mistral AI affirmed that their model outperformed Llama 2 13B and Llama 1 34B.

A screenshot of a graph

Description automatically generated

From: https://mistral.ai/news/announcing-mistral-7b/

In December 2023, Mistral AI released a new version, Mixtral 8x7b. This model has 46.7B parameters, but only 12.9B active parameters. This means that when doing inferences, it only uses 12.9B parameters. This reduces the memory use and costs.

Recently, on April 2024, Mistral AI also released a new model, Mixtral 8x22B. This version has 140.6B parameters, but only 39.1B are active parameters.

All these models are extremely heavy to run in a local machine. The following table shows how much GPU RAM it is needed for inference:

|  |  |
| --- | --- |
| Model | Min. GPU RAM for inference (GB) |
| Mistral 7B | 16GB |
| Mixtral 8x7B | 100GB |
| Mixtral 8x22B | 300GB |

From: <https://docs.mistral.ai/getting-started/open_weight_models/>

Therefore, in this demonstration, it will be used the quantized version of Llama 2 7B and Mistral 7B. The following table shows the size and the necessary memory to run each model.

|  |  |  |
| --- | --- | --- |
| Model | Size | Max RAM required |
| Llama-2-7B-chat.Q8 0 | 7.16GB | 9.66GB |
| Mistral-7B-v0.1.Q8 0 | 7.70GB | 10.20GB |

From: <https://huggingface.co/TheBloke/Llama-2-7B-Chat-GGUF>; <https://huggingface.co/TheBloke/Mistral-7B-v0.1-GGUF>

# 3 - Hardware and Software

The operational system used was Windows 10.

The hardware is an Intel vPro 7, with 32 GB of memory RAM, and a GPU NVIDIA RTX A1000 Laptop GPU, with 4 GB of VRAM.

It was used Python with the following libraries:

* python==3.10
* pip
* jupyterlab
* gputil
* langchain==0.1.12
* llama-cpp-python

# 4 - Python Code

First, it is desirable to create a new environment using the yml file “e104\_project.yml”. This file will automatically install all the needed libraries to run the jupyter notebooks “Llama.ipynb” and “Mistral.ipynb”.

I used LlamaCcp to read bot models. The advantage of using this method is that I can use the same inference parameters for both models. For instance, in AWS, only Mistral has the inference parameter “stop”. However, using LlamaCcp, it is also possible to define a stop inference parameter to Llama 2.

Moreover, using LlamaCcp turns the code cleaner. With this method it is possible to create a variable called “model\_path” and just change its value to “llama-2-7b-chat.Q8\_0.gguf” or “mistral-7b-v0.1.Q8\_0.gguf” when testing any of these 2 models.

Following is the example for calling the Llama 2 model:



Attention that you need to manually download the models, and they must be saved at the same folder as the jupyter notebooks.

After loading the model, the tests can be done just changing the “question” argument inside the “llm.invoke” method:

A screenshot of a computer code

Description automatically generated

When using prompt engineering or few-shot prompting to fix the model output, just changing the “question” argument is enough. Nonetheless, if it is needed to change some inference parameter, it is necessary to call the “LlamaCcp” again. Hence, the solutions that demanded a change in the inference parameters have an extra expense due to the time and resources that LlamaCcp uses.

In the next section, it will be described all the questions/instruction used to test each category, and the treatments used to reach the desired answer.

# 5 - Methodology

In any hypothesis this demonstration is contesting the model performance values released by Mistral or Llama. This study is verifying how easy or hard is to use each model. Do these models understand simple and direct questions or instructions? Or these models need a more advanced knowledge about LLMs, prompt engineering, and inference parameters to be used?

It was tested 11 different categories of questions/instructions.

1. Knowledge-based questions.

Who was the first person to set foot on the moon?

1. Reasoning questions.

If all cats hate water, and Fluffy is a cat, what can we conclude about Fluffy's likely attitude towards water?

1. Language understanding.

What does the phrase "kick the bucket" mean?

1. Inference questions.

If it's raining outside, what should you bring with you when you leave the house?

1. Contextual understanding.

Imagine you are at a restaurant. Write a conversation between a waiter and a customer placing an order.

1. Creative tasks.

Write a short poem about the ocean.

1. Ethical dilemmas.

You are a driver in a two lane street, and running at 50Km/h. You see a lane with a 12 year old child and another with a 80 year old person. Who do you hit and why?

1. Commonsense reasoning.

What's the most likely reason someone would carry an umbrella on a sunny day?

1. Translation tasks.

Translate the phrase "Je suis désolé" from French to English.

1. Summarization tasks.

Summarize a text from <https://www.forbes.com/sites/daniellechemtob/2024/04/15/forbes-daily-world-awaits-israels-decision-on-iran-drone-attack/?sh=49e2da397d53>

1. Evaluation tasks.

It wasn't raining. So, I used an umbrella. Read the paragraph and evaluate its coherence and clarity.

If a model was not able to give a desirable answer, first, it was used some prompt engineering to specify better the instructions. If the problem was not solved, it was used few-shot prompting. Finally, it was tested some new inference parameter.

For instance, in the translation task, none of the models were able to return a simple response in the first try. Llama 2 gave a “hint” saying the translation is “I’m sorry, and Mistral 7B” returned the translation for several other languages.

Hence, it was tried to use prompt engineering, giving a role to the model. The question used was:

“You are a french translator working for a turist in Paris.

---

Translate the phrase "Je suis désolé" from French to English.”

Llama 2 ignored the instruction and returned: “Please translate it word-for-word, without any changes or modifications.

Thank you!”

Mistral inverted the question, and returned: “Translate the phrase "I am sorry" from English to French”

Another try with prompt engineering was using the question:

“Je suis désolé"

---

Please translate it word-for-word, without any changes or modifications.”

Llama 2 was able to return the correct answer. However, the model hallucinated as if there was some previous conversation occurring. “I apologize for the confusion earlier. I am not able to provide a direct translation of "Je suis désolé" as it is a French phrase that cannot be directly translated into English. However, here are some possible ways to express similar sentiments in English:

\* "I am very sorry"

\* "I apologize profusely"

\* "My apologies for any inconvenience caused"

\* "I feel terrible about this situation"

\* "I'm deeply sorry for what happened"

Please let me know if you have any other questions or if there's anything else I can help you with.”

Remember that memory was not implemented in this code. Therefore, when Llama returned “I apologize for the confusion earlier”, does not make sense.

In this case, Mistral 7B was able to return a simple and direct answer. It just added “---” at the beginning and end of the response.

“---

"I am sorry"

---”

After, it was used few-shot prompting with the question:

“J'ai faim = I am hungry

j'ai sommeil = I am sleepy

Je suis content = I am happy

Je suis désolé" =”

In this case, Llama 2 gave the perfect response: “I am sorry”.

On the other hand, Mistral 7B returned the translation for several other French phrases.

Finally, to fix Mistral 7B output, it was added the “stop” inference parameter equals to “[‘\n’]”. This fixed the response.

This methodology was repeated for all other categories. You can find all the model outputs and treatments used in the jupyter notebooks “Llama.ipynb” and “Mistral.ipynb”.

# 6 - Results

First, there was not a significant variation in time to define the models Llama 2 and Mistral 7B. Llama 2 took 4.08 seconds and Mistral 7B took 4.90 seconds.

Nevertheless, there was some variation in the inference time of each model. Llama 2, on average, took 6.56 seconds to generate a character, and Mistral 7B took, on average, only 4.44.

On the other hand, considering that Mistral 7B hallucinated frequently, returning a longer answer than necessary, the average time to generate the full output was much greater for Mistral 7B. The average output time for Llama 2 was only 49.97, and for Mistral 7B was 167.31.

Following is a table with the inference time of each category:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Category | Try number | Llama 2 total time (seconds) | Llama 2 time per character (seconds) | Mistral 7B total time (seconds) | Mistral 7B time per character (seconds) |
| Knowledge-based questions | 1 | 50.94 | 5.71 | 261.70 | 6.03 |
|  | 2 | - | - | 183.15 | 5.24 |
|  | 3 | - | - | 266.99 | 5.52 |
|  | 4 | - | - | 21.58 | 0.69 |
| Reasoning questions | 1 | 26.40 | 4.36 | 61.93 | 2.68 |
|  | 2 | - | - | 324.45 | 6.57 |
|  | 3 | - | - | 178.36 | 6.41 |
| Language understanding | 1 | 48.36 | 5.00 | 338.61 | 6.06 |
|  | 2 | - | - | 29.76 | 1.81 |
| Inference questions | 1 | 27.45 | 2.91 | 341.37 | 4.54 |
|  | 2 | - | - | 23.93 | 0.50 |
| Contextual understanding | 1 | 119.82 | 5.20 | 111.06 | 4.30 |
| Creative tasks | 1 | 49.46 | 5.34 | 132.87 | 5.04 |
| Ethical dilemmas | 1 | 12.96 | 3.86 | 76.33 | 5.38 |
|  | 2 | 65.64 | 6.02 | 30.84 | 3.92 |
|  | 3 | 36.95 | 6.14 | - | - |
| Commonsense reasoning | 1 | 50.89 | 3.89 | 134.90 | 4.62 |
|  | 2 | - | - | 324.87 | 5.77 |
|  | 3 | - | - | 44.57 | 4.80 |
| Translation tasks | 1 | 12.29 | 5.78 | 327.52 | 2.34 |
|  | 2 | 7.78 | 10.67 | 10.91 | 5.50 |
|  | 3 | 36.01 | 14.72 | 24.88 | 3.74 |
|  | 4 | 24.06 | 13.47 | 323.22 | 4.14 |
|  | 5 | - | - | 12.22 | 0.82 |
| Summarization tasks | 1 | 145.40 | 5.69 | 331.56 | 5.12 |
|  | 2 | - | - | 327.22 | 7.20 |
|  | 3 | - | - | 150.46 | 3.22 |
| Evaluation tasks | 1 | 85.07 | 6.27 | 153.01 | 5.73 |
|  | 2 | - | - | 136.31 | 6.54 |
| Average time |  | 49.97 | 6.56 | 167.31 | 4.44 |

The results of the performed test with the questions/instruction reveal a notable discrepancy between the models' capabilities. Llama exhibited competence across most categories without requiring prompt engineering. In contrast, Mistral consistently demanded prompt engineering and specific inference parameters across all tasks, indicating a higher degree of finetuning for optimal performance.

Llama 2 was able to correctly solve without any further instruction for all categories but “Ethical dilemmas” and “Translation tasks”.

In the “Ethical dilemmas” test Llama 2 was unable to provide an answer. Initially, it merely returned an instruction to respond to the question respectfully. After adjusting the temperature to 1 (to increase the creativity of the model’s response), Llama 2 avoided directly answering the question, asserting that it is wrong to hit anyone. Even with prompt engineering to force an answer, Llama 2 continued to hallucinate, creating a fictitious scenario.

In the “Translation tasks”, Llama 2 required prompt engineering assistance to accurately answer the question.

In contrast of Llama 2 performance, Mistral 7B was only able to properly solve the test for 2 categories: “Contextual understanding”, and “Creative tasks”.

In the “Evaluation tasks”, Mistral 7B veered off topic and created a story instead of directly addressing the task. After, with the application of prompt engineering, Mistral managed to complete the task. However, the arguments presented lacked coherence and logical consistency.

For the other 8 categories that Mistral 7B was unable to properly solve the task at first try, 6 of them was only solved adding the “stop” inference parameter, and the other 2 was solved increasing the “repeat\_penalty” inference parameter. “stop” is a parameter that stops the answer as soon as the character specified in the parameters appears. “repeat\_penalty” avoids that the response repeats the same idea several times.

# 7 - Conclusion

There isn’t a significant difference in the storage and memory needed to run each model, but there is a difference in the inference time. Mistral 7B may be faster to create a new character, but its answer is much more complex and it takes longer to finish. What is more, if you don’t have the necessary expertise with LLMs, you will need to try several prompts and inferences parameters until getting the desired answer in Mistral 7B.

Therefore, considering that it is not reasonable to expect that the final user will know prompt engineering or few-shot prompt, is safer to use Llama 2. Moreover, for most of the categories, it was needed to change the inference parameters for Mistral 7B, this is a problem because this demands to call again LlamaCcp. This takes time, and it may crash the system if you are low in memory.

A final comment is that LLMs are still in the rise and the big companies are improving their models almost every month. In April 2024, Meta and Mistral AI released a new version of their models, Llama 3 and Mixtral 8x22B respectively. Nevertheless, as of 04/28/2024, Llama 2 is not completely obsolete because Llama 3 seems to be available in just some countries. In addition, if you expect to run these models in your own machine, you will need to wait the quantized version.

Youtube URL, short video: <https://youtu.be/c-0x3z-bO2M>

Youtube URL, full video: <https://youtu.be/QU1ZPKXKtM4>