**INFO 290**

**Assignment:** LLM Comparative Evaluation Using Ollama

**Name:** Javad Aminian-Dehkordi

**Computer Used to run tests (Model, Memory, GPUs?) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**LLM Models (name and size)**

Select three models that represent different size categories to compare performance across the spectrum.

**NOTE:** *I had issues with running the large models. In this assignment, I had two small model* (TinyLlama/TinyLlama-1.1B-Chat-v1.0 *and* microsoft/phi-2) *and a medium one* (mistralai/Mistral-7B-Instruct-v0.1).

* **First LLM: TinyLlama/TinyLlama-1.1B-Chat-v1.0**
* **Second LLM: microsoft/phi-2**
* **Thirrd LLM: EleutherAI/gpt-neo-1.3B**

**Prompts**

**Prompt P1 (Factual Knowledge):**

* **Explain the concept of overfitting in machine learning to a high school student.**

**Prompt P2 (Instruction Following with Constraints):**

* **Write a short story involving a robot learning to feel emotions for the first time.**

**Measure Prompts on each Model**

**Detailed Evaluation Breakdown**

**Evaluation Scale for Accuracy:**

* **5**: Completely accurate with precise details
* **4**: Mostly accurate with very minor errors
* **3**: Generally accurate but with some notable errors
* **2**: Partially accurate with significant errors
* **1**: Mostly inaccurate information

**[First LLM] Evaluation**

| **Prompt** | **Accuracy (1-5)** | **Time to complete** | **Prompt Tokens** | **Response Tokens** | **Response Time** |
| --- | --- | --- | --- | --- | --- |
| Factual Knowledge |  | <20 sec |  |  |  |
| Instruction Following |  | <20 sec |  |  |  |

**[2nd LLM] Evaluation**

| **Prompt** | **Accuracy (1-5)** | **Time to complete** | **Prompt Tokens** | **Response Tokens** | **Response Time** |
| --- | --- | --- | --- | --- | --- |
| Factual Knowledge |  | ~ 2 min |  |  |  |
| Instruction Following |  | ~ 2 min |  |  |  |

**[3rd LLM] Evaluation**

| **Prompt** | **Accuracy (1-5)** | **Time to complete** | **Prompt Tokens** | **Response Tokens** | **Response Time** |
| --- | --- | --- | --- | --- | --- |
| Factual Knowledge |  |  |  |  |  |
| Instruction Following |  |  |  |  |  |

**Prompt Evaluation Summary**

| **Model Name** | **Size** | **P1 Accuracy** | **P2 Accuracy** | **Avg Accuracy** | **Avg. Tokens** | **Avg. Time** |
| --- | --- | --- | --- | --- | --- | --- |
| [1st LLM] |  |  |  |  |  |  |
| [2nd LLM] |  |  |  |  |  |  |
| [3rd LLM] |  |  |  |  |  |  |

**Key Observations and Analysis**

**Summarize your key findings comparing the three models**

**How would you describe the relationship between Model Size and Performance?**

**What are some potential use cases for different model sizes.**

**Reflection**

What surprised you most about the differences between the models?

What challenges did you encounter when running the models locally?

How might these findings influence your choice of LLM for different applications?

**What did you learn from the LLM generated code?**

The code provided to run the LLM from its API was written with the assistance of Claude. When using existing code and/or LLM generated code, take it as a learning opportunity.

Select a line or multiple lines from the provided code such that when you first read it you said “Hmm, not sure what that does” or “I didn’t know you could do that ..” . Perhaps there were no surprises. That’s OK too. Just report “No Insights”

The Code: <provide here>

Do a bit of research and describe what you learned.