**Assignment No: 4**

**Name : Aakash Sukre**

**Student ID : 23200858**

**Link :** [**https://chainforge.ai/play/?f=ihtezqecivxt**](https://chainforge.ai/play/?f=ihtezqecivxt)Use this link to see the flow and implementation.

**Aim:**

Here, in this assignment we discuss the results of our empirical assessment that tested how using Chain of Thought (CoT) prompts affected the logical reasoning abilities of different Language Model Models (LLMs). Eight examples encompassing a range of logic and math issues with difficulty levels from moderate to hard were chosen using the GSM8K dataset.

The study employed the following methodology:

**Data Source:** The GSM8K dataset, containing various difficulty levels of math problems, was used (<https://huggingface.co/datasets/gsm8k>)

**LLM Selection:** The trial started with Falcon.7B, Mistral.7B. Instruct because of initial financial constraints. GPT-3.5.

**ChainForge Flow Design:**

* Load test challenges, a flow was created in ChainForge.
* I used two prompt nodes:

1. One that specifically guides the LLM through the reasoning process with CoT prompts.
2. One that presents the issue head-on without using CoT.

* Go through the selected LLMs using the prompts.
* Compare the LLM responses to the anticipated responses.
* Make a visual representation of the success rates.

**Methodology:**

Data Preparation: A sample of problems that is indicative of the GSM8K dataset was chosen by downloading it. Across difficulty levels (easy, medium, hard) and issue categories (arithmetic, logic, word problems), a balanced distribution was sought in the selection process.

**System Architecture:  
  
Chain Forge Flow Design:**

Tabular Data Node: Consist of 8 questions with expected answer.

**A screenshot of a chat

Description automatically generated**

**Prompt Nodes:**

1. Falcon-7B
2. Mistral-7B
3. GPT3.5

I have given prompt here –

Help me to solve the {Question}

* 1. **CoT Prompt Node:**
  2. A screenshot of a computer

     Description automatically generatedAlex has 10 pencils. He goes to the store and buys 4 packs of pencils. Each pack contains 5 pencils. How many pencils does he have now?
  3. Answer:

Alex starts with 10 pencils. Four packs of 5 pencils each is 20 pencils. 10 + 20 = 30 so the answer is 30 pencils.

**Non-CoT Prompt Node:**

* 1. Alex has 10 pencils. He goes to the store and buys 4 packs of pencils. Each pack contains 5 pencils. How many pencils does he have now?
  2. answer is 30 pencils.

**LLM Nodes:** Sent the prompts to the respective LLMs – Falcon-7B, Mistral-7B, GPT3.5

**Vis Node:** Presented the success rates of each LLM with and without CoT for all test problems.

**A screenshot of a computer

Description automatically generated**

**Evaluation Results:**

**Falcon.7B:** At first, Falcon-7B's success rate was moderate, demonstrating its aptitude for solving simple arithmetic and logic problems. Its performance differed amongst examples, though, suggesting that its capacity for reasoning is limited. Falcon.7B showed an improvement in accuracy with CoT prompts, especially when explicit reasoning was required.

**Mistral.7B:** Mistral.7B demonstrated a somewhat greater success rate in contrast to Falcon.7B, suggesting a superior understanding of rational thinking. Like Falcon.7B, though, its performance differed amongst samples. Mistral.7B also demonstrated progress with CoT prompts, particularly in sequential reasoning scenarios.

**GPT-3.5:** In every scenario, GPT-3.5 consistently outperformed Falcon.7B and Mistral.7B, with a 100% success rate. It was clear that it could produce well-reasoned answers even in the absence of clear instructions. But the inclusion of CoT prompts improved its answers even more, demonstrating a more coherent argument.

**Analysis of the Findings:**

It is possible to attribute the observed enhancement in LLM performance with CoT prompts to the following factors:

Structured Reasoning: By offering a methodical approach, CoT prompts help the LLM deconstruct the issue and proceed with a reasonable and systematic cognitive process.

Clear Directions: Emphasizing crucial phases in the reasoning process could assist the LLM in keeping their attention on the important components of the issue and preventing them from becoming distracted by other details.

**There are restrictions on this study:**

Restricted Scope: A subset of the GSM8K dataset's problems and a limited number of LLMs—originally Falcon.7B.Instruct were I used in this experiment.

Cost Restrictions: Falcon.7B.Instruct was first chosen since it was free to use.

A screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generated

In this above Image - Throughout my evaluation, I noted that Chain-of-Thought (CoT) prompting has the potential to improve the reasoning capabilities of Language Model Models (LLMs), with models such as Falcon.7B, Mistral.7B and GPT 3.5. I was able to fully evaluate the effect of CoT prompts on LLM performance over a range of logical and mathematical reasoning tasks by using a varied problem set taken from the GSM8K dataset. Furthermore, I emphasized the value of error analysis in pinpointing the places where CoT dramatically enhances performance, offering insightful information about the advantages and disadvantages of various LLMs. We learn more about LLM capabilities and gain insight into how well CoT prompting works to improve logical reasoning abilities when these factors are taken into account throughout the evaluation process.

A screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generated

From the above chart - I submitted eight questions for the experiment, ranging in difficulty from easy to medium to sophisticated. The results revealed a noteworthy pattern for these questions: although GPT-3.5 consistently produced the most accurate answers, Falcon.7B and Mistral showed lower accuracy rates. While Mistral performed marginally better than Falcon, neither model was very good at giving precise responses, which frequently led to false alarms. These results demonstrate GPT-3.5's superior performance and point out the difficulties Falcon.7B and Mistral encountered when trying to solve the provided issues. Although Mistral performed marginally better than Falcon, both models had trouble producing acceptable results, suggesting that their reasoning skills needed to be strengthened.

In order to thoroughly evaluate the reasoning skills of Language Model Models (LLMs), I used eight different difficulty-level questions from the GSM8K dataset. Calculating Layla's part of the apple picking, deciding on complimentary ice cream cones based on sales revenue at Dan's store, and comparing the amount of trout caught by Caleb's dad to Caleb were among the tasks. The logical and mathematical thinking abilities of the LLMs were put to the test in several settings.

**Conclusion:**

To sum up, this practical assessment provided insightful information on how well Chain-of-Thought (CoT) prompting supports Language Model Models' (LLMs') reasoning abilities. I found that GPT-3.5 consistently outperformed Falcon.7B and Mistral in a variety of models, demonstrating its skill at solving logical and mathematical reasoning problems. Even while Mistral and Falcon performed somewhat better than each other, both models had flaws that frequently led to incorrect answers. This emphasizes how important it is to keep improving and adjusting LLMs in order to improve their capacity for reasoning. Furthermore, the application of a heterogeneous problem set and error analysis was essential in pinpointing the domains in which CoT prompting markedly improved performance. All things considered, my research adds to the current effort to improve LLM capabilities by emphasizing the possibility of CoT prompting as a useful tactic for enhancing logical reasoning skills in language models.