

LLMs for Educating Hateful meme creators

-By Devarasetty Sri Vaibhav

<https://arxiv.org/pdf/2210.03629.pdf>

Problem Statement Goals:-

Target Audience: Individuals involved in creating or sharing memes with harmful content related to nationality, race, religion, gender, etc.

Goal: To educate about the impact of hateful memes while promoting empathy, understanding, and responsible online behavior.

Problem Statement

- Making an LLMs which provides education to the people who do hateful memes(based on nationality, race, religion, gender, etc.)

Example:-

Topic :- Dehumanising Muslim Religion

Education :- In today's diverse and interconnected world, it is essential that we foster understanding and respect for all religions and cultures. This includes the Muslim religion, which is one of the world's major faiths. Unfortunately, there are instances where individuals or groups dehumanize Muslims, perpetuating stereotypes and misconceptions. Education is a powerful tool to combat such prejudice and promote empathy and inclusivity. Ways to educate people and encourage a more respectful and informed perspective on the Muslim religion.

1)Understand the Basics 2)Combating Stereotypes of Muslims 3)Interfaith Dialogue 4)Diverse Perspectives 5)Community Engagement Etc.

Approaches for the PS

- We have to make a dataset which contains hateful memes and the Appropriate Educative Content. This is LLM would also train upon the `attack_type`(dehumanising, mocking, inferiority, slurs, etc.) and the `target_group`(nationality, religion, race, gender, etc.), these attributes would anchor our LLM to provide us with better educative content for the hateful memes creators.
- Then we train many LLMs model with the above crafted dataset which has the educative content for each type of hateful memes.
- We then test our all of our LLM models with LLM benchmarks to evaluate our model's performance.

Approaches for making the Dataset

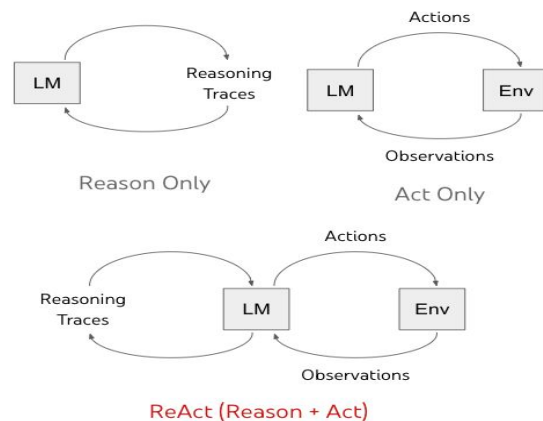
- 1) Manual LLM Prompts :- Asking LLMs to provide appropriate prompts to ask Wiki Search Engine, and then Summarizing the top wiki results for providing education.

These could give proper results, but it involves a lot of manual human efforts to get the proper dataset and we the results we get can be also not appropriate for the Education LLM we would make, and also we have to run the prompts, multiple times if they are any changes in the information.

- 2) LLM Agents :- These LLM Agents are just RL based LLM agents which provide better prompts for getting better data points for our dataset, instead of straight up sending the prompts for getting information, it uses a prompting algorithm to get the information we need.

ReAct - SYNERGIZING REASONING AND ACTING IN LANGUAGE MODELS

- ReAct is a new concept that combines reasoning and acting in language models.
- The Framework enables language models to generate verbal reasoning traces and text actions.
- These actions interact with external environments, while reasoning traces update the model's internal state.
- The alternation between reasoning and acting is based on task requirements.



Flow of the Framework

Thought: ...

Action: ...

Observation: ...

... (Repeated many times)

Actions -> Environment Query Processing -> Observations from the environment -> Use Reasonable Tracing for create the next action the LLM should take -> Actions

Actions of the LLMs is the query we send to the environment (ex:- Wiki, Knowledge Bases, Etc.)

Observations are the output of the query we get from the environment.

There are various types of useful reasoning traces, e.g., decomposing task goals to create action plans, injecting commonsense knowledge relevant to task solving, extracting important parts from observations, tracking task progress while maintaining plan execution, handling exceptions by adjusting action plans, and so on.

ReAct Prompting

(1) Hotspot QA

Question: How many rooms are in the hotel that is home to the Cirque du Soleil show Mystere?

HotspotQA label: 2,664

Outdated ✗

(1a) Standard

Answer: 3,000 ✗

(1b) CoT (Reason Only)

Thought: Let's think step by step. The hotel that is home to the Cirque du Soleil show Mystere is Treasure Island. Treasure Island has 2,885 rooms, so the answer is 2,885.

Answer: 2,885 ✗

(1c) Act-Only

Act 1: Search[Cirque du Soleil]

Obs 1: Cirque du Soleil is a Canadian entertainment company...

Act 2: Lookup[Mystere]

Obs 2: No more results

Act 3: Search[Mystere (Cirque du Soleil show)]

Obs 3: Could not find...

Act 4: Search[Mystere (Cirque du Soleil)]

Obs 4: Mystère (my-steer) is one of six...

Act 5: Lookup[Treasure Island Hotel and Casino]

.....(end without answer) ✗

(1d) ReAct (Reason + Act)

Thought 1: I need to search Cirque du Soleil show Mystere, find the hotel it is in, then find the number of rooms in the hotel.

Act 1: Search[Cirque du Soleil show Mystere]

Obs 1: Could not find Cirque du Soleil show Mystere.

Similar: ['Cirque du Soleil' ...]

Thought 2: To find the hotel, I can search Mystere (Cirque du Soleil).

Act 2: Search[Mystere (Cirque du Soleil)]

Obs 2: Mystère (my-steer) is ... Held in a custom theatre at the Treasure Island Hotel and Casino ...

Thought 3: Mystere is held in a custom theatre at the Treasure Island Hotel and Casino. So I need to search Treasure Island Hotel and Casino next and find the number of rooms in the hotel.

Act 3: Search[Treasure Island Hotel and Casino]

Obs 3: Treasure Island Hotel and Casino ... with 2,884 rooms and 220 suites ...

Thought 4: Treasure Island Hotel and Casino has 2,884 rooms and 220 suites. So the answer is 3,104.

Act 4: Finish[3,104]

Up-to-date ✓

Results of ReAct

The approach with the best results is a combination of ReAct and CoT that uses both internal knowledge and externally obtained information during reasoning.

	HotpotQA (exact match, 6-shot)	FEVER (accuracy, 3-shot)
Standard	28.7	57.1
Reason-only (CoT)	29.4	56.3
Act-only	25.7	58.9
ReAct	27.4	60.9
Best ReAct + CoT Method	35.1	64.6
Supervised SoTA	67.5 (using ~140k samples)	89.5 (using ~90k samples)

PaLM-540B prompting results on HotpotQA and Fever.

Progress

- We have figured out the approach of the problem statement, and have found LLM Agents to be useful for curating the dataset.
- Have brought some few example prompting outputs from the Langchain Based LLM agent
- We have setup some LLM models(example:- LLAMA2, GPT3) to use for the React Based-prompting to get the dataset for the Problem statement.