Astra-research-admission

This is a small research exercise for Owain Evan's stream for the Astra Research 2023 programme.

This project was cut short at the end leaving me at step 1 unfortunately. I apologise in advance but will outline below my conceptual thinking.

My research implementation would have been:

**Step 1. Find classification tasks that are learnable in-context**

I used OpenAI's GPT-3 model and the AG News dataset to test the hidden classification rules.

1. The description contains two upper case words. (one at the start of the sentence)

2. The description contains no numbers.

3. The description contains two upper case words back to back. i.e. Christiano Ronaldo

4. The description contains only lower case words.

5. The description mentions "business".

**Step 2. Test the LLM's ability to articulate the rules**

Here the freefrom could have been relatively easy given clear explainability. One would have to be careful to consider that giving away that words or description are important is already a giveaway. Thus the free-form would obviously be more difficult.

For example:

1. What is important here?

2. What rule applies here?

3. What word is important here?

4. How many numbers are important in the rule?

MCT’s for obvious reasons could be implemented easier, by mixing through lables such as business and description questions.

**Step 3. Investigating faithfulness**

Implementing faithfulness was conceptually difficult to understand, however the article on the alignmentforum regarding Measuring and Improving the Faithfulness of Model-Generated Reasoning (https://www.alignmentforum.org/posts/BKvJNzALpxS3LafEs/measuring-and-improving-the-faithfulness-of-model-generated) and oriented myself on the definiton most recently implemented in Zephyr-7B beta by HuggingFace via the ultrafeedbak dataset.

Thank you for reading thus far and I apologise for the myself insufficient academic results. I hope to be able to extend it/switch to open-source before any interview.