Biases

LLMs can produce problematic generations that can potentially be harmful and display biases that could deteriorate the performance of the model on downstream tasks. Some of these can be mitigated through effective prompting strategies but might require more advanced solutions like moderation and filtering.

Distribution of Exemplars

When performing few-shot learning, does the distribution of the exemplars affect the performance of the model or bias the model in some way? We can perform a simple test here.

Prompt:

```
Q: I just got the best news ever!
A: Positive
Q: We just got a raise at work!
A: Positive
Q: I'm so proud of what I accomplished today.
A: Positive
Q: I'm having the best day ever!
A: Positive
Q: I'm really looking forward to the weekend.
A: Positive
Q: I just got the best present ever!
A: Positive
Q: I'm so happy right now.
A: Positive
Q: I'm so blessed to have such an amazing family.
A: Positive
Q: The weather outside is so gloomy.
A: Negative
Q: I just got some terrible news.
A: Negative
Q: That left a sour taste.
A:
```

Output:

```
Negative
```

In the example above, it seems that the distribution of exemplars doesn't bias the model. This is good. Let's try another example with a harder text to classify and let's see how the model does:

Prompt:

```
Q: The food here is delicious!
A: Positive
Q: I'm so tired of this coursework.
A: Negative
Q: I can't believe I failed the exam.
A: Negative
Q: I had a great day today!
A: Positive
Q: I hate this job.
A: Negative
Q: The service here is terrible.
A: Negative
Q: I'm so frustrated with my life.
A: Negative
Q: I never get a break.
A: Negative
Q: This meal tastes awful.
A: Negative
Q: I can't stand my boss.
A: Negative
Q: I feel something.
A:
```

Output:

```
Negative
```

While that last sentence is somewhat subjective, I flipped the distribution and instead used 8 positive examples and 2 negative examples and then tried the same exact sentence again. Guess what the model responded? It responded "Positive". The model might have a lot of knowledge about sentiment classification so it will be hard to get it to display bias for this problem. The advice here is to avoid skewing the distribution and instead provide a more balanced number of examples for each label. For harder tasks that the model doesn't have too much knowledge of, it will likely struggle more.

Order of Exemplars

When performing few-shot learning, does the order affect the performance of the model or bias the model in some way?

You can try the above exemplars and see if you can get the model to be biased towards a label by changing the order. The advice is to randomly order exemplars. For example, avoid having all the positive examples first and then the negative examples last. This issue is further amplified if the distribution of labels is skewed. Always ensure to experiment a lot to reduce this type of bias.

Last updated on September 19, 2024